Protein Binding Site Comparison

The Impact of Binding Site Similarity

on Hit Identification in Early Drug Discovery

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften

(*Doctor rerum naturalium*)

an der

Fakultät für Chemie und Chemische Biologie

Technische Universität Dortmund

vorgelegt von

M. Sc.

Christiane Ehrt

aus Zeitz

Dortmund, 2019

Erstgutachter: Prof. Dr. Daniel Rauh

Zweitgutachter: Prof. Dr. Stefan M. Kast

Die vorliegende Arbeit wurde in der Arbeitsgruppe von Dr. Oliver Koch unter der Leitung von Prof. Dr. Daniel Rauh in der Zeit von Januar 2014 bis April 2019 am Lehrstuhl für Medizinische Chemie an der Fakultät Chemie und Chemische Biologie der Technischen Universität Dortmund angefertigt.

KOOPERATIONEN

Einige Ergebnisse dieser Dissertation sind in enger Zusammenarbeit mit Herrn Dipl.-Inf. Tobias Brinkjost entstanden, der maßgeblich für die Programmierung der Methode zur Klassifikation von Sekundärstrukturelementen (SCOT) und der Software zum Vergleich der räumlichen Anordnung von Helices und Strands auf der Grundlage von Graphen (SLOT) verantwortlich war. Beide Methoden wurden im Arbeitskreis gemeinsam entwickelt.

PUBLIKATIONSLISTE

Zhao, L.; Ehrt, C.; Koch, O.; Wu, Y.-W. One-pot N2C/C2C/N2N ligation to trap weak protein-protein interactions. *Angew. Chem. Int. Ed. Engl.* **2016**, *55* (28), 8129–8133.

Mejuch, T.; Garivet, G.; Hofer, W.; Kaiser, N.; Fansa, E. K.; Ehrt, C.; Koch, O.; Baumann, M.; Ziegler, S.; Wittinghofer, A.; Waldmann, H. Small-molecule inhibition of the UNC119-cargo interaction. *Angew. Chem. Int. Ed. Engl.* **2017**, *56* (22), 6181–6186.

Spinka, M.; Seiferheld, S.; Zimmermann, P.; Bergner, E.; Blume, A.-K.; Schierhorn, A.; Reichenbach, T.; Pertermann, R.; Ehrt, C.; König, S. Significance of individual residues at the regulatory site of yeast pyruvate decarboxylase for allosteric substrate activation. *Biochemistry* **2017**, *56* (9), 1285–1298.

Kaitsiotou, H.; Keul, M.; Hardick, J.; Mühlenberg, T.; Ketzer, J.; Ehrt, C.; Krüll, J.; Medda, F.; Koch, O.; Giordanetto, F.; Bauer, S.; Rauh, D. Inhibitors to overcome secondary mutations in the stem cell factor receptor KIT. *J. Med. Chem.* **2017**, *60* (21), 8801–8815.

Maurer, S.; Buchmuller, B.; Ehrt, C.; Jasper, J.; Koch, O.; Summerer, D. Overcoming conservation in TALE-DNA interactions: a minimal repeat scaffold enables selective recognition of an oxidized 5-methylcytosine. *Chem. Sci.* **2018**, *9* (36), 7247–7252.

Ehrt, C.; Brinkjost, T.; Koch, O. Binding site comparison – Software and applications. *Encyclopedia of Bioinformatics and Computational Biology*; Elsevier, 2019; pp 650–660.

Im Rahmen dieser Dissertation sind außerdem die folgenden Publikationen entstanden, die in Teilen in das vorliegende Manuskript eingeflossen sind:

Ehrt, C.; Brinkjost, T.; Koch, O. Impact of binding site comparisons on medicinal chemistry and rational molecular design. *J. Med. Chem.* **2016**, *59* (9), 4121–4151.

Ehrt, C.; Brinkjost, T.; Koch, O. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLoS Comput. Biol.* **2018**, *14* (11), e1006483.

Satisfaction of one's curiosity

is one of the greatest sources

of happiness in life.

Linus Pauling

TABLE OF CONTENTS

ZUSAMMENFASSUNG 1

ABSTRACT 3

ABBREVIATIONS 5

1 INTRODUCTION 7

1.1 Binding Site Comparison in Medicinal Chemistry 8

1.1.1 Small Molecule Binding Sites 8

1.1.2 The Identification of Binding Sites 8

1.1.3 The Comparison of Ligand Binding Sites 11

1.2 Binding Site Comparison Based on Secondary Structure Elements 13

1.2.1 Secondary Structure Elements in Proteins 13

1.2.2 The Comparison of the Spatial Arrangement of Secondary Structure Elements 15

1.3 Trypanothione Synthetase 17

1.3.1 Neglected Tropical Diseases 17

1.3.2 Trypanothione Synthetase as a Promising Drug Target 18

1.3.3 Trypanothione Synthetase and Protein Kinases 19

2 AIM OF THE THESIS 23

3 METHODS 25

3.1 Binding Site Prediction 25

3.1.1 Analysed Standalone Software for Binding Site Prediction 25

3.1.2 Dataset Preparation 27

3.1.3 Binding Site Identification - Necessary Pre-Processing Steps and Usage Details 28

3.1.4 Evaluation of the Prediction Performance 35

3.1.5 Establishment of a Combined Approach to Improve Prediction Performance 38

3.2 Binding Site Comparison 40

3.2.1 Dataset Preparation 40

3.2.2 Dataset Characterisation 40

3.2.3 Datasets 40

3.2.4 Binding Site Comparison 43

3.2.5 Performance Assessment 46

3.2.6 Runtime Analyses 48

3.3 Secondary Structure Assignment 49
3.3.1 Clustering of Turns According to Their Conformation 49
3.3.2 Analysis of the Turn-Based Secondary Structure Classification Method SCOT 58
3.4 Trypanothione Synthetase 59
3.4.1 Comparative Modelling 59
3.4.2 MD Simulations 59
3.4.3 Paullone Binding Mode Prediction 60
3.4.4 Comparison to Protein Kinases Based on the Secondary Structure Elements 61
3.4.5 Binding Site Comparison 64
3.4.6 Docking of the Protein Kinase Similarity-Based Compound Library 65
4 RESULTS AND DISCUSSION 67
4.1 Binding Site Prediction 67
4.1.1 Motivation 67
4.1.2 Binding Site Prediction Methodologies 67
4.1.3 Identification of Appropriate Benchmark Datasets 83
4.1.4 Challenges for Binding Site Prediction Benchmark Studies 88
4.1.5 Quality Criteria for the Evaluation of Binding Site Prediction Methodologies 88
4.1.6 The Extended PDBbind Core Set 90
4.1.7 The Dataset of Cryptic Binding Sites 102
4.1.8 Preliminary Conclusions 111
4.1.9 Binding Site Prediction for Homology Models and Allosteric Sites 113
4.1.10 Establishment of a Meta Approach to Improve Prediction Performance 119
4.1.11 Conclusion 122
4.2 Binding Site Comparison 124
4.2.1 Motivation 124
4.2.2 Choice of Binding Site Comparison Tools for Evaluation 124
4.2.3 Binding Site Comparison Methodologies 126
4.2.4 Benchmark Datasets 132
4.2.5 Benchmark Studies 134
4.2.6 Can We Distinguish Between Similar and Dissimilar Sites? 160
4.2.7 Runtime Analyses 164
4.2.8 Conclusion 166

4.3 Secondary Structure Assignment 171
4.3.1 Motivation 171
4.3.2 Turn Classification 171
4.3.3 Assignment of Helices and Sheets with SCOT 214
4.3.4 Conclusion 220
4.4 A Case Study: Trypanothione Synthetase 222
4.4.1 Motivation 222
4.4.2 Comparative Modelling and MD Simulation Studies 222
4.4.3 Druggability Assessment 224
4.4.4 The Paullone Binding Mode 225
4.4.5 The Initial Search for TryS Inhibitors 227
4.4.5 TryS and Protein Kinases – A Common Ligand-Sensing Core? 228
4.4.6 TryS and Protein Kinases – Similarities in the ATP Binding Site Properties 232
4.4.7 Drug Repurposing and Related Strategies 234
4.4.8 Conclusion 237
CONCLUSION AND OUTLOOK 239
SCHLUSSFOLGERUNG UND AUSBLICK 243
REFERENCES 247
APPENDIX 1: Binding Site Prediction A1.1
APPENDIX 2: Binding Site Comparison A2.1
APPENDIX 3: Secondary Structure Assignment A3.1
APPENDIX 4: Trypanothione Synthetase A4.1
ACKNOWLEDGEMENTS
Eidesstattliche Versicherung (Affidavit)

ZUSAMMENFASSUNG

Ziel der vorliegenden Arbeit war die Identifikation neuer Hitverbindungen zur Inhibition der Trypanothionsynthetase (TryS) in Organismen der Gattung Leishmania. Diese Parasiten sind die Auslöser für einige der sogenannten vernachlässigten Krankheiten. Neue Therapeutika zur Behandlung dieser Erkrankungen sind dringend notwendig, da die gegenwärtigen Behandlungsmethoden – im Vergleich zu jenen für typische Erkrankungen in Industrieländern – nicht hinreichend verlässlich sind.

Der grundsätzliche Ansatz dieser Arbeit war die Etablierung eines strukturbasierten Workflows mit Fokus auf Ähnlichkeiten zwischen den ATP-Bindetaschen von TryS aus *Leishmania major* und Proteinkinasen. Letztere sind gut charakterisierte Enzyme und die geplante Herangehensweise sollte vom umfangreichen Wissen zu wirkstoffähnlichen Verbindungen zur Inhibition von Proteinkinasen profitieren.

Allerdings erfordert die Etablierung einer ähnlichkeitsbasierten Suche nach möglichen Hits das harmonische Zusammenspiel zahlreicher Programme des molekularen Modellierens. Der aufgebaute Workflow umfasst die Homologiemodellierung der Targets, MD Simulationen, eine umfassende Analyse der Bindetascheneigenschaften, den Vergleich der Bindetaschen von TryS und Proteinkinasen sowie ein nachfolgendes Docking bekannter Kinaseinhibitoren. Dazu wurden Inhibitoren der Proteinkinasen, deren ATP-Bindetaschen die stärksten Gemeinsamkeiten mit denen des Targets aufweisen, selektiert.

Die rasanten Entwicklungen auf dem Gebiet der *in silico* Hitidentifizierung führten zum Aufkommen zahlreicher Methoden zur Lösung verschiedenster Aufgabenstellungen. Bestimmte Softwaretools auf dem neuesten Stand der Technik haben sich für regelmäßig angewandte Ansätze (molekulares Docking, Pharmakophorsuchen) durchgesetzt. Für weniger häufig eingesetzte Methoden zu Vorhersage und Vergleich von Proteinbindetaschen ist dies jedoch nicht der Fall. Eine Vielzahl parallel entwickelter Tools mit grundsätzlich gleichen Zielstellungen, aber unterschiedlichen Eigenschaften war die Folge. Diese Diversität erfordert eine zuverlässige Gegenüberstellung der Methoden zur Herausarbeitung der geeignetsten Software zur Bearbeitung bestimmter wissenschaftlicher Fragestellungen. Der Hauptteil dieser Arbeit widmet sich daher der umfassenden Analyse von Software zur Vorhersage und zum Vergleich von Proteinbindetaschen, um deren Stärken und Schwächen kritisch gegenüberzustellen. Die Ergebnisse dieser Evaluationen bieten mögliche Kriterien für die Wahl einer passenden Methode. In Bezug auf die Bindetaschenvorhersage war es möglich, außergewöhnlich zuverlässige Methoden aufzuzeigen. Anders ist es im Hinblick auf den automatisierten Vergleich von Bindetaschen. In diesem Fall ist eine exakte Einschätzung der Voraussetzungen und Erfordernisse des Projekts unerlässlich und ein Konsensusansatz scheint in einigen Fällen die angemessenste Vorgehensweise zu sein.

Abgesehen von Bindetaschenähnlichkeiten auf Grundlage der Form und physikochemischen Eigenschaften, wurde ein weiterer alternativer Ansatz untersucht. Basierend auf der Erkenntnis, dass entfernt verwandte Proteine identische molekulare Grundgerüste binden und über eine ähnliche geometrische Organisation von Sekundärstrukturelementen im Bereich der Bindetasche (l*igand-sensing cores*) verfügen, wurde vorgeschlagen, dass ein Ansatz zum Auffinden ähnlicher Sekundärstrukturanordnungen in den Bindetaschen bekannter Proteinstrukturen Hinweise zur Identifikation neuer Binder für noch weitestgehend unerforschte Targets liefern könnte. Dieser Ansatz wurde als möglicher Anhaltspunkt für die Suche nach ähnlichen Liganden von Proteinkinasen und TryS verfolgt. Allerdings war es im Rahmen dieser Arbeit nicht möglich, eine erfolgreiche Strategie zu etablieren.

Der Erfolg in der Entwicklung einer neuen Methodik zur Identifikation aussagekräftiger Gemeinsamkeiten hinsichtlich der Sekundärstrukturelementanordnungen in unmittelbarer Umgebung der Liganden hängt maßgeblich von den zur Verfügung stehenden Evaluierungsdatensätzen und der Zuverlässigkeit des verwendeten Ansatzes zur Zuordnung von Sekundärstrukturelementen ab. Erst wenn beide Voraussetzungen angemessen erfüllt werden können, ist es möglich, die Anwendbarkeit des Konzepts der *ligand-sensing cores* an interessanten Targets zu verifizieren. Mit der Entwicklung der Methode SCOT zur Klassifikation von Sekundärstrukturelementen konnte die erste Hürde in dieser Hinsicht genommen werden. Ihre Stärken liegen in der geometrischen Regelmäßigkeit und Stabilität der zugeordneten Sekundärstrukturelemente. Dennoch wird die Optimierung einer Methodik zum Vergleich von Sekundärstrukturanordnungen durch das Fehlen geeigneter Datensätze weiterhin erschwert.

Bezüglich des Targets dieser Studie konnte hingegen gezeigt werden, dass etablierte Methoden zum Vergleich von Proteinbindetaschen hervorragend geeignet sind, Ähnlichkeiten zwischen den Nukleotidbindetaschen von TryS und Proteinkinasen aufzuzeigen. Bekannte Inhibitoren der Kinasen mit den signifikantesten Ähnlichkeiten zum Target sowie chemisch ähnliche Verbindungen wurden untersucht. Abschließend konnten Vorschläge für vielversprechende Moleküle zur Adressierung der ATP-Bindetasche von TryS gemacht werden, deren Potenzial zur Inhibition von TryS aus Leishmania nun anhand von biochemischen Testungen belegt werden muss.

ABSTRACT

The aim of the present study was the identification of novel hit compounds for the inhibition of trypanothione synthetase (TryS) from organisms of the genus Leishmania. They are the causative agents of so-called neglected diseases. There is an urgent need for novel therapeutics for these diseases as the current treatment methods are not sufficiently reliable when compared to the medical management for diseases prevalent in developed high-income countries.

The basic idea of this work was to set up a structure-based drug repurposing workflow based on similarities between the ATP binding sites of TryS from *Leishmania major* and protein kinases. The latter enzymes are well characterised, and this approach benefits from the huge knowledge of drug-like small molecule inhibitors of protein kinases.

However, the establishment of such a workflow requires the reliable interplay of multiple software methods for *in silico* structure-based modelling. The chosen workflow includes the comparative modelling of the enzymes under investigation, MD simulation studies, an in-depth analysis of pocket properties, the comparison of binding sites of TryS and kinases, and a subsequent molecular docking of known inhibitors of protein kinases which are most similar to TryS with respect to the ATP binding site properties.

The breakneck developments in the field of *in silico* hit identification led to the emergence of a myriad of available methods to meet similar targets. Current state-of-the-art methods evolved for often applied structure-based methods such as molecular docking or pharmacophore searches. Unfortunately, methods for the prediction, analysis, and comparison of binding sites are less frequently applied. This led to the parallel development of numerous methods which differ in multiple aspects. This diversity necessitates reliable comparison and benchmark analyses to elucidate the most suitable approaches for a pre-defined scientific question. In consequence, a main part of the present work is dedicated to the comprehensive evaluation of developed software for the prediction and comparison of binding sites. The analyses shed light on the major strengths and restrictions of the methods under investigation. The results ensure that only the most reliable methods are chosen for the problem under investigation. For binding site prediction methods, we can pinpoint a restricted set of exceptionally reliable methods. The case is different for binding site comparison tools. We must finally conclude that the establishment of a consensus approach might be the most appropriate way to identify binding site similarities.

Besides the comparison of binding sites based on physicochemical and shape properties, an alternative approach was tested for its applicability for the presented project. Based on the observation that some less closely related proteins binding to similar scaffolds share a common arrangement of SSEs in the ligand binding sites, it was hypothesised that finding similar arrangements of SSEs in unrelated proteins might facilitate the search for novel promising scaffold for yet unexplored enzymes. This so-called “ligand-sensing cores” approach was adopted for the comparison of ATP binding sites of TryS and kinases.

However, the results for this case study indicate not only that it is difficult to pinpoint specific and unique secondary structure element arrangements in protein binding sites, but also that the success in establishing a novel method rises and falls with the availability of proper optimisation datasets and the reliability of secondary structure assignment methods. We tried to solve at least the latter part of the problem by developing the novel secondary structure assignment tool SCOT which classifies geometrically uniform helices and strands. But especially with respect to parameter optimisation, further developments of SSE comparison methods and the generation of suitable datasets will be necessary.

However, with respect to the target of this study, we could show that well-established methods for the comparison of protein binding sites are perfectly suitable to pinpoint similarities between the nucleotide binding sites of TryS and protein kinases. Known inhibitors of protein kinases with the most significant similarities to the target and structurally related compounds were investigated. Finally, we could propose promising molecules for addressing the ATP binding site of TryS which must now be evaluated experimentally for their potential to inhibit TryS from Leishmania.

ABBREVIATIONS

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
APT1	acyl-protein thioesterase 1
ATP	adenosine triphosphate
AUC	area under the ROC curve
CCDC	Cambridge Crystallographic Data Centre
CDK2	cyclin-dependent kinase 2
Cf	*Crithidia fasciculata*
CI	confidence interval
CL	cutaneous leishmaniasis
CSD	Cambridge Structural Database
DSSP	Define Secondary Structure of Proteins
Ec	*Escherichia coli*
EC	Enzyme Commission
EF	enrichment factor
ESOM	Emergent Self-Organizing Map
GO	Gene Ontology
GOLD	Genetic Optimization for Ligand Docking
GSH	glutathione
GSK-3β	glycogen synthase kinase-3β
GSP	glutathionyl spermidine
GspS	glutathionyl spermidine synthetase
GUI	graphical user interface
HAT	human African trypanosomiasis
HSSP	Homology-derived StructureS of Proteins
KLIFS	Kinase-Ligand Interaction Fingerprints and Structures
LBVS	ligand-based virtual screening
Li	*Leishmania infantum*
Lm	*Leishmania major*
LSD1	lysine-specific demethylase 1
MAO	monoamine oxidase
MAPK	mitogen-activated protein kinases
MCC	Matthews correlation coefficient
MCS	maximum common subgraph

MD	molecular dynamics
MIF	molecular interaction field
MSA	multiple sequence alignment
NMR	nuclear magnetic resonance
NTD	neglected tropical disease
PCA	principle component analysis
PDB	Protein DataBank
PI4K2A	phosphatidylinositol 4-kinase type IIα
PKACA	cAMP-dependent protein kinase catalytic subunit α
PKDL	post-kala-azar dermal leishmaniasis
PLD	Protein Ligand Database
POI	protein of interest
PPI	protein-protein interaction
PPII	polyproline II
PSSC	Protein Structure Similarity Clusters
PSSM	Position-Specific Scoring Matrix
RF	Random Forest
RMSD	root-mean-square deviation
RMSF	root-mean-square fluctuation
ROC	receiver operating characteristics
ROCS	Rapid Overlay of Chemical Structures
SAH	S-adenosyl-L-homocysteine
SAS	solvent accessible surface
SBVS	structure-based virtual screening
SHAFT	Secbase Automated Helix Assignment adapted from Turns
SSE	secondary structure element
SVM	Support Vector Machine
Tb	*Trypanosoma brucei brucei*
Tc	*Trypanosoma cruzi*
TryS	trypanothione synthetase
$T(SH)_2$	trypanothione
vdW	van der Waals
VL	visceral leishmaniasis
VS	virtual screening

1 INTRODUCTION

According to the WHO, neglected tropical diseases (NTDs) affect approximately one billion people in 149 countries worldwide.[1] Several NTDs are caused by protozoan organisms of the genus *Trypanosoma* and *Leishmania*. Current treatment methods are prone to severe side effects, low efficacy, and high costs.[2] Therefore, the search for suitable small molecular drugs to cope with diseases like human African trypanosomiasis (HAT) is an ongoing topic. Trypanothione synthetase (TryS), an enzyme involved in the redox metabolism of the parasites, is the model target of these studies. In this work, it was attempted to answer the question of whether the binding site analysis and comparison can facilitate the search for novel hit compounds inhibiting TryS.

Various available tools to predict, characterise, and compare small molecule binding sites were investigated to identify their individual strengths and weaknesses. A special focus was on the analysis of the impact of secondary structure information on site comparisons. Based on the known similarities between TryS and protein kinases with respect to the molecular scaffolds of the bioactive compounds, the similarities and differences between the adenosine triphosphate (ATP) binding sites of both enzymes were investigated in more detail.

Finally, a drug repurposing workflow was developed considering the detected similarities between two protein kinases and TryS. The basic challenge of the study was the huge number of available methods which can be applied. Given this and the restricted number of benchmark datasets and studies, this work investigates whether we can rationally decide on the methods which are most suitable for the problems to be addressed.

1.1 Binding Site Comparison in Medicinal Chemistry

1.1.1 Small Molecule Binding Sites

The in-depth analysis of small molecule binding sites in proteins is one of the crucial steps for rational drug design. It is one of the major determinants for the success or failure of various *in silico* strategies to find and develop small molecule modulators of protein activity. Typical structure-based virtual screening (SBVS) approaches include pharmacophore searches, molecular docking, and various methodologies to estimate the free energy of small molecule binding to macromolecules.[3] These methods cannot only support the progress of the identification of novel lead structures,[4] but can also assist in the improvement of purely ligand-based virtual screening (LBVS) studies.[5] Starting at an early stage of *in silico* drug discovery, the identification of promising targets and the elucidation of druggable binding sites require the application of different computational methods before further steps are taken. In some cases, the 3D structure of the protein of interest (POI) is unknown. Comparative modelling can help to establish a structural model of the protein. This process requires the identification of related structures with reasonably high sequence identity. Knowledge of (evolutionary) relationships additionally helps to unravel putative functions and cellular locations.[6] Sometimes, the structure of the POI is already published, but the location of active and regulatory sites is still unknown. Due to the high number of published structural models of proteins, which is rapidly increasing in the Protein DataBank (PDB),[7] a structure comparison might support the elucidation of the respective cavities. Nevertheless, even in the absence of such similarities, computational tools were developed to predict potential small molecule binding sites.

Some methods also provide measures for the druggability of the proposed sites,[8] i.e., the potential of a site to accommodate drug-like compounds leading to an effect on the protein function.[9] The abundance of not only biologically interesting small molecules in the PDB, but also buffer ions, organic solvents, and modified protein residues renders it a challenging task to search for druggable binding sites in all publicly available protein models. To solve this problem, the sc-PDB[10] offers a set of annotated druggable binding sites for further analyses.

1.1.2 The Identification of Binding Sites

The elucidation of ligand binding sites in proteins is an important starting point for various SBVS projects. The annotation of previously unknown binding sites and subsequent comparison to known sites can help to disclose distinct relationships and can inspire the identification of novel targets. Or the other way around, it can facilitate the search for the binding site of interest of an experimentally identified target and assist in target validation. Tools for the automated prediction of putative small molecule binding sites can be classified according to the basic underlying assumptions. Geometry-based methods solely rely on the cavity shape features such as depth, buriedness, and volume (e.g., PASS[11] and LIGSITE^{CS}[12]), whereas energy-based methods predict putative sites of small molecules by considering the physicochemical nature of the solvent accessible protein residues (e.g., GRID[13]) or make use of molecular docking approaches (e.g., BindingResponse[14]).

The observation that the "binding site space" is degenerate with respect to shape and chemical properties lays the cornerstone for conservation-based methods.[15] A comparison of sequences and identification of conserved residues based on different similarity measures enables us to infer function and detect potential binding sites with methods such as ConSurf.[16] The same holds true for a comparison of complete protein structures or binding sites as shown with PDBinder.[17] These tools are summarised using the term knowledge-based methodologies.

The differences in the nature of the binding sites hamper the development of a universally applicable, high-performance method to predict all types of relevant cavities. The high flexibility of binding sites as observed in induced fit and conformational selection phenomena is another challenge when using a single structure of the POI. Furthermore, cryptic sites which can often only be detected in the presence of a ligand or with the help of MD (molecular dynamics) simulations demand a more elaborate analysis of the protein.[18] Although conservation-based methods might also achieve promising results for this cavity type, a complete disregard of protein flexibility will often lead to dissatisfying results in such cases. Binding site similarity-based methods will only detect potential binding sites if a similar cavity is already known. Additionally, the shape of the binding site can have a huge impact on the prediction outcome. Geometry-based methods might fail in the case of buried or narrow binding sites. Different tools are available that are specially developed for the detection of interior voids (e.g., BetaVoid[19]) or restricted to enclosed pockets (e.g., VOIDOO[20]). Special cases of interest are flat protein-protein interfaces, huge solvent-exposed nucleic acid binding sites, or small phosphate binding sites. They gave rise to the development of highly non-generic methods that are only applicable to a small subset of cavity types.

Examples are ProMate[21] and DoBi[22] for protein-protein interfaces, BindUP[23] and DBS-PRED[24] for DNA or RNA binding sites, Nucleos[25] and NSiteMatch[26] for nucleotide binding sites, IonCom[27] for ion binding sites, MetalDetector[28] and CHED[29] for metal binding sites, and Phosfinder[30] for phosphate binding sites. These tools might become interesting for highly specific structure-based projects, while their suitability, e.g., for function prediction or protein annotation, is limited. Additionally, the elucidation of allosteric binding sites has come to the fore over the last years. Examples for methods trying to detect sites of allosteric regulation are SPACER,[31] AlloPred,[32] and AllositePro.[33] Researchers who are interested in certain binding site types might rely on those methods, especially as protein-protein interfaces or carbohydrate binding sites are typical cases where early binding site identification approaches failed.[18,34] However, the elucidation of small molecule binding sites, in general, necessitates the use of methods with a very broad applicability domain.

Geometry and energy-based methods (especially approaches that utilise blind docking experiments) highly rely on the side-chain conformations of the pocket residues, while knowledge-based methods (based on the residue conservation or binding site similarities) will fail for novel and unique structures. The introduction of meta prediction tools successfully addressed the problem of the individual strengths and drawbacks of the detection methods.

A combination of the predictions of different tools and a final rescoring with approaches such as MetaPocket[35] or COACH[36] leads to improved performance and higher prediction accuracy. Moreover, elaborate workflows enable a reliable prediction of putative cavities of interest.[37] Nonetheless, they are not suitable for large-scale analyses of available protein databases and restricted to a small number of POIs.

A further challenge is the ranking of the identified protein cavities. Although it was shown that the cavities with the largest volume often constitute true binding sites,[38] the results of different methods for various datasets revealed the drawbacks of this ranking scheme.[39] Therefore, the importance of druggability predictions or machine learning-based pocket ranking and rescoring schemes has grown over the last years.[40,41]

Finally, the high number of available standalone tools and web servers for the prediction of binding sites poses an additional challenge. The decision for a promising tool is thwarted by the fact that there is no gold standard for the evaluation of small molecule binding site prediction software. A large-scale benchmark study is still owing. It is of interest to find a combination of complementary methods that yield a convincing performance in a meta approach. Figure 1.1.1 illustrates the cumulative frequency of developed approaches from the 1990s onwards. The fact that this frequency reaches a plateau suggests that we can now fall back upon sufficiently reliable methods to predict putative small molecule binding sites. Nonetheless, there is no extensive knowledge as to which of the tools performs best for a given dataset. The available benchmark studies[42,43] and reviews[37,44–47] do not cover a broad range of methods. Consequently, the individual strengths and weaknesses of the available methods remain largely unknown. Finally, missing knowledge of reliable tools leads to general problems for the development and evaluation of novel methods.

In general, it becomes infeasible to evaluate all available tools against a new one. This leads to the in many cases random selection of tools which are nowadays used for comparison purposes, including tools which are not publicly available anymore.

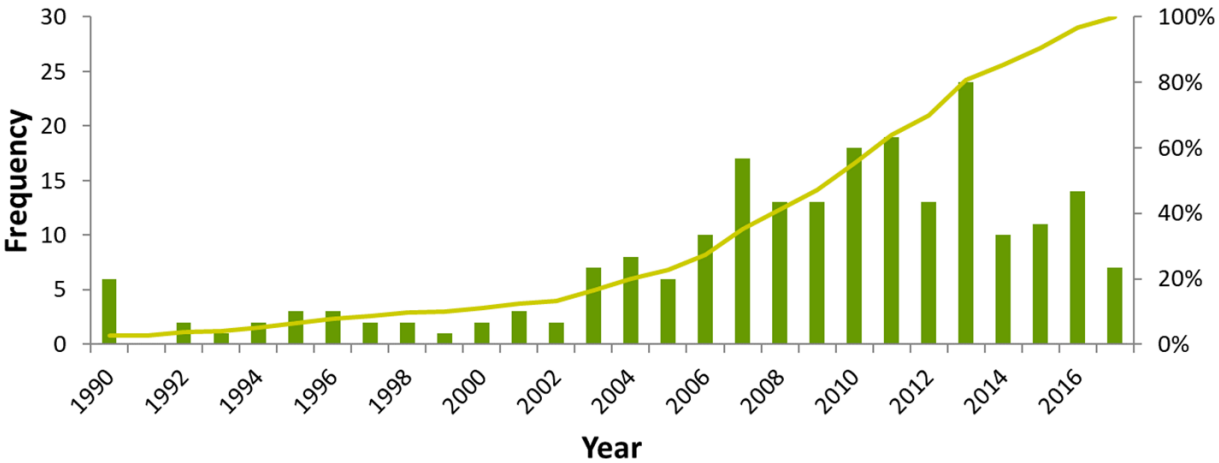


Figure 1.1.1. Developments in the automated prediction of small molecule binding sites in proteins. Frequencies (dark green bars) and cumulative frequencies (light green line) of publications concerning novel or improved binding site prediction methods from 1990 until 2017 are presented. A complete list is given in Table A1.1 in Appendix 1.

1.1.3 The Comparison of Ligand Binding Sites

The comparison of ligand binding sites has proven beneficial for various drug discovery projects.[48] Due to the ever increasing number of publicly available protein structures in the PDB,[7] binding site comparison has become a promising tool to exploit accessible knowledge and transfer it to other scientific projects. Together with the structural data, the number of available binding site comparison methods has considerably grown over the last years. Therefore, it becomes increasingly difficult to decide on a suitable method for a certain field of research. Figure 1.1.2 presents different application domains of binding site comparison studies and tools that have been successfully applied for these particular challenges. While some methods (e.g., PocketMatch[49]) provided novel insights for various research scenarios, the use of other tools (e.g., VolSite/Shaper[50]) was restricted to a more specific focus. In consequence, it is difficult to derive suggestions for promising tools based on the results presented in the literature. Although a number of review papers highlight the impact of binding site comparison on structure-based target assessment[51] and pharmaceutical discovery,[52] the number of benchmark studies is notably small, as is the number of suitable datasets for the comprehensive evaluation of the available methods. Therefore, the second part of the analyses presented herein is intended to derive guidelines which facilitate the choice of an appropriate tool for comparing ligand binding sites. This was achieved by exhaustively evaluating a selected set of successfully applied methods. A unique benchmark analysis was performed by designing independent and objective datasets that can elucidate the individual strengths of the tools.

However, it must be stated that binding site similarity lies in the eye of the beholder. Therefore, the applicability domains of a method that was designed to compare protein-ligand interaction patterns will be different from those of a comparison approach which relies on binding site residues or surface properties. Another challenge is the appropriate ranking of cavity similarities.[53] In this context, it is even more difficult to differentiate between similar and dissimilar binding sites. The development of a consensus method that considers the output of multiple binding site comparison software might enable a more reliable differentiation.

Given these yet ambiguous definitions of binding site similarity, it becomes difficult to investigate phenomena such as promiscuity by means of binding site comparisons. In the most general definition, promiscuity is the binding of a ligand to multiple, not necessarily related targets (ligand promiscuity) or, vice versa, the binding of a protein to a broad spectrum of ligands (protein promiscuity). This phenomenon can be exploited to establish polypharmacology or to find small molecule modulators of novel targets by means of drug repurposing strategies. On the other hand, ligand promiscuity is also the reason for the binding of drugs to off-targets which might lead to undesirable side effects. As long as the molecular reasons for promiscuity are unknown, it is difficult to investigate how to utilise the potential of promiscuity for the design of drugs, but avoid the obvious drawbacks of off-target effects. As shown before,[54,55] binding site similarities offer explanations for promiscuity. However, the question persists on how to reliably determine this similarity.

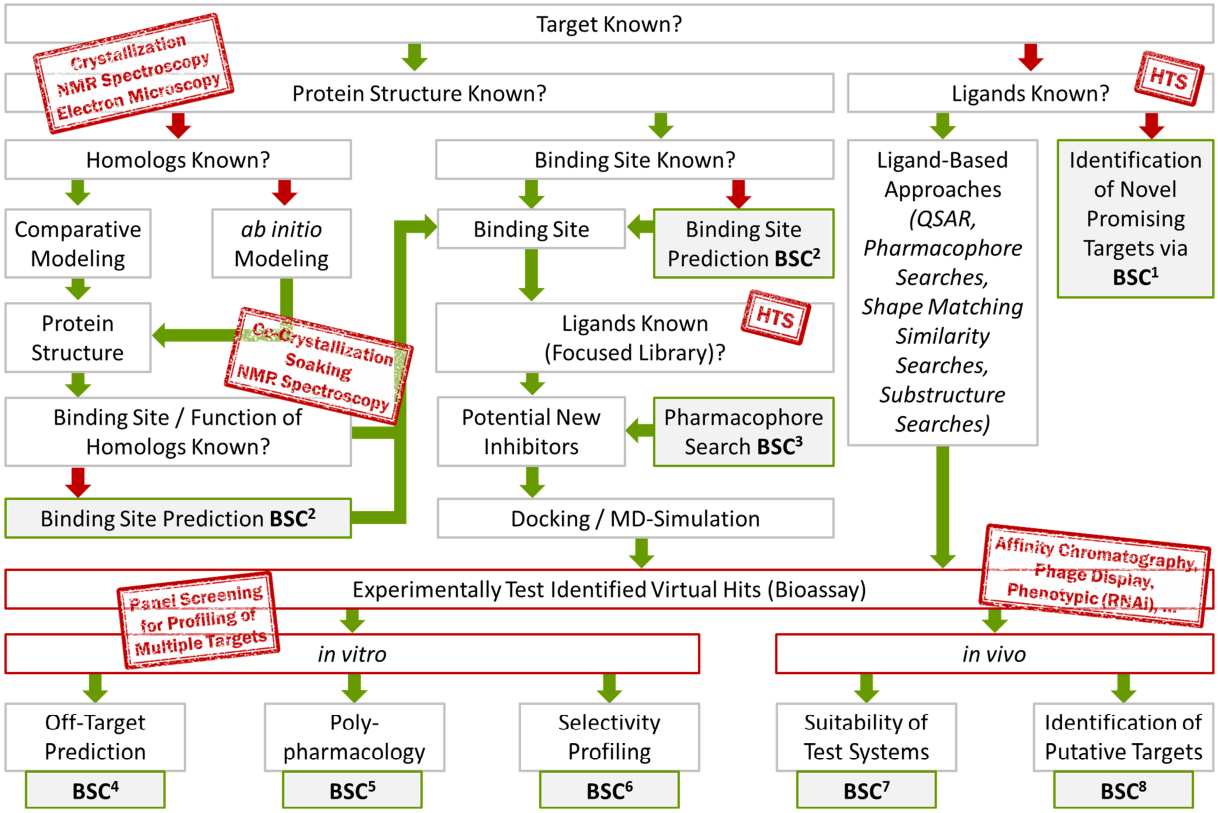


Figure 1.1.2. Workflow illustrating different stages of structure-based *in silico* drug design that can be supported by the application of binding site comparison methods. Alternative biochemical approaches to the computational methods are given in red. BSC[1] refers to the identification of novel promising targets as realised with the comparison Method of Wu *et al.*[56] and IsoCleft Finder.[57] Binding site comparison also enables the identification of unknown binding sites which was achieved using COFACTOR,[58] eF-site,[59] PocketMatch,[49] PoLiMorph,[60] ProBiS,[61] SiteEngine,[62] and SurfaceScreen[63] (BSC[2]). Drug repurposing strategies were followed by similarity analyses using Cavbase,[64,65] IFP,[66] CPASS,[67] PSSC,[68] SMAP,[69] SOIPPA,[70] and TM-align[71] (BSC[3]). The prediction of putative off-targets was achieved with APF,[72] FuzCav,[73] KRIPO,[74] the Method of Brakoulias and Jackson,[75] ProBiS,[61] PSIM,[76] Shaper,[50] SOIPPA,[70] and SiteAlign[77] (BSC[4]). Further application domains are polypharmacology analyses (BSC[5], as shown for FuzCav,[73] PocketFEATURE,[78] PocketMatch,[49] Shaper,[50] and SMAP[69] analyses), the prediction of selectivity profiles (BSC[6], using APF,[72] Cavbase,[64,65] FLAP,[79] the Method of Brakoulias and Jackson,[75] and PESD[80]), the assessment of test system suitability (BSC[7], using APF,[72] CMASA,[81] and PocketMatch[49]), and the identification of promising targets (BSC[8], using KRIPO[74]).

1.2 Binding Site Comparison Based on Secondary Structure Elements

1.2.1 Secondary Structure Elements in Proteins

Protein structures can be characterised by four hierarchical levels starting from the primary structure (the amino acid sequence) and secondary structure (distinct repetitive backbone conformations of structure fragments) leading towards the tertiary (the protein domains as independent folding units) and quaternary structure (the assembly of protein chains to form a polymeric functional unit). Pauling *et al.* were the first researchers to discover regular secondary structure elements (SSEs) as a crucial level in this protein structure hierarchy. They analysed the underlying interactions and the geometry of the protein building blocks referred to as helices and strands.[82] Their stability is maintained by distinct hydrogen bonding patterns between the protein backbone atoms that also determine the class of the corresponding SSE.

Although these patterns are by no means regular and bifurcated or even missing hydrogen bonds are observed, the geometry of the SSEs is characterised by distinct ranges of the backbone φ and ψ angles. Ramachandran plots allow a crude assignment of the protein residues to the respective SSE.[83] While helices are composed of consecutive internal hydrogen bonding patterns, β-strands stabilise each other in parallel and antiparallel β-sheets. These sheets consist of two or more β-strands which are hydrogen-bonded in an antiparallel, parallel, or mixed fashion. Due to these interaction patterns, the residues within SSEs show distinct φ and ψ angles. Figure 1.2.1 gives a schematic overview of the hydrogen bonding patterns and dihedral angles observed for commonly occurring helix and β-strand classes.

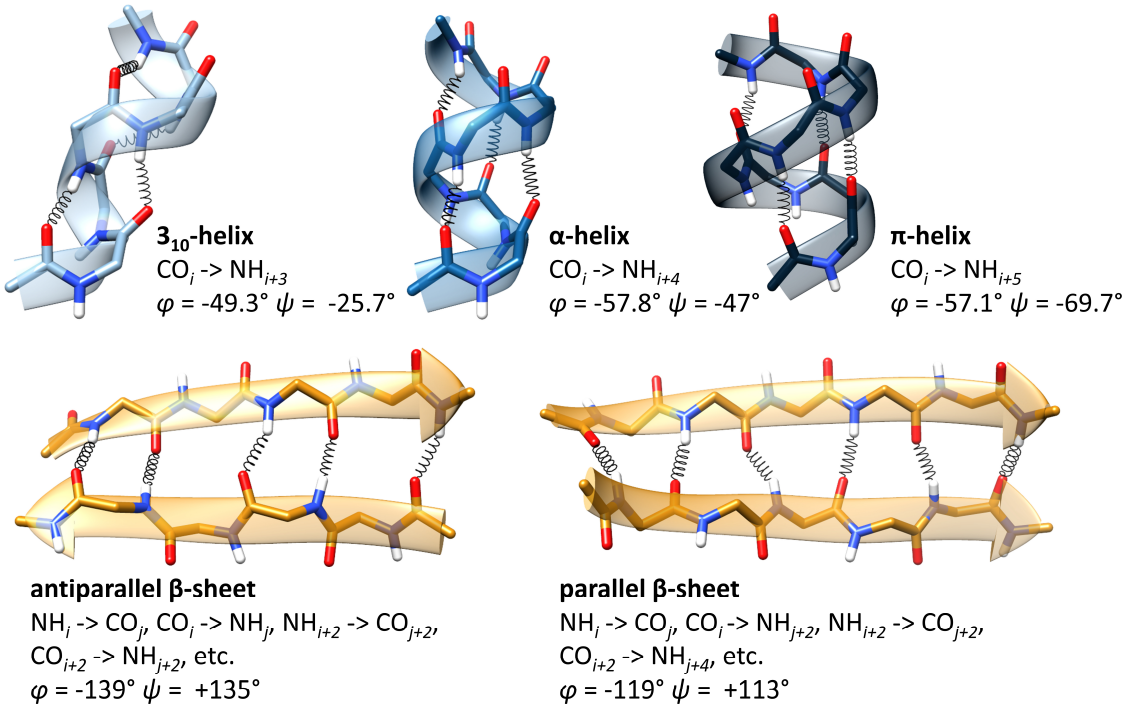


Figure 1.2.1. Showcase structures for the most commonly observed SSEs in proteins. Top: The three most common right-handed helix types in protein structures with their respective hydrogen bonding pattern (black springs) and proposed dihedral angles.[84–86] Bottom: Hydrogen-bonding patterns (black springs) in antiparallel and parallel β-sheet structures. The corresponding proposed dihedral angles for both types are also given.[87,88]

In addition, we can define PPII (polyproline II) helices[89] and left-handed helices[90] which are rarely observed in protein structures. Their existence was even questioned.[91] However, some functional roles of these SSE types are known.[90,92] In this context, regular backbone conformations are not only crucial for the structural integrity of proteins, but also for enzyme catalysis and regulation.[93]

Of all known helix types, the α-helix is the most prominent one when analysing all publicly available protein structure models. This helix is not only crucial for protein folding and stability. Its helix dipole enables crucial functionalities of helices in binding, catalysis, and regulation.[94,95] Cooley *et al.* very comprehensively discussed the origin and function of π-helices in proteins which are underrepresented in protein structures. The small number of analysed π-helices can be partially attributed to a hierarchy which is commonly applied in helix type assignment.[96] Usually, these helices coexist with α-helices (α-bulges), are derived from single residue insertions, and seem to be conserved within proteins.[97] In contrast, Enkhbayar *et al.* discussed the occurrence of 3_{10}-helices, which are irregular in terms of the hydrogen bond pattern and their geometry within proteins.[98] The presence of these helices in potassium channels and membrane proteins hint at their potential regulatory function.[99]

The functional roles of β-sheets are diverse. They stabilise folds which are essential for protein functionality, e.g., the β-barrel structure in green fluorescent protein which influences the chromophore nature of the protein.[100] For ATP grasp enzymes, a β-sheet is crucial to shield the active site from the solvent which prevents the auto-hydrolysis of the bound nucleotide.[101] In other proteins, they act as the link between monomers of multimeric functional units or even as the regulator of multimer formation.[102] Additionally, β-sheet formation plays an undesired role in prion-related diseases.[103] Knowledge of the different protein-folding diseases underpins the crucial meaning of SSEs for protein function.

Numerous methods have been developed to assign the SSEs of the 3D structures of proteins. Table 1.2.1 gives an overview of the methods, their availability, and their individual approaches. Most methods rely on geometric features (e.g., dihedral angles, Cα positions, etc.), but the still most commonly used methods DSSP (Define Secondary Structure of Proteins)[104] and STRIDE (STRuctural IDEntification)[105] are mainly based on the backbone hydrogen bonds. Given the plethora of available methods and the fact that the majority can be readily used, the question arises whether the scientific community benefits from yet another approach. Intriguingly, most of the recently developed methods are geometry-based, whereas the frequently cited and applied tools rely on hydrogen bond criteria.

Besides SHAFT (Secbase Automated Helix Assignment adapted from Turns),[106] none of the available methods classifies SSEs using dihedral angle-derived turn type assignments. The basic drawback of this classification is the missing assignment of strands and kink residues in unusually bent regions of helices and strands. This encouraged us to develop a novel turn-based classification method which assigns the common helix types, β-sheet structures, and rare helix types, such as PPII and left-handed helices. The problems arising from the application of a helix class hierarchy were also addressed.

The resulting method, SCOT (Secondary structure Classification On Turns) additionally enables the geometry-based identification of kinks in helices and strands leading to geometrically consistent SSE assignments.

Table 1.2.1. Secondary structure assignment methods. The methods and the basic underlying approaches are given. Software which is available (either on request, as a downloadable tool, or as a web server) for the processing of PDB files is given in italics.

Hydrogen Bonds	Dihedral Angles	Backbone Geometry	Method
yes			*SECSTR*,[107] *PROMOTIF*,[108] *SSTRUC*,[109,110] *DSSP*,[104] *DSSPcont*[111]
yes	yes		*PSSC*,[112] *SHAFT*,[106] *DSSP-PPII*,[113] *STRIDE*[105]
yes		yes	Method of Chen *et al.*,[114] Method of Levitt and Greer,[115] β-Spider[116]
yes	yes	yes	SCOT
	yes		*DISICL*,[83] *PROSS*[117]
	yes	yes	*KAKSI*,[118] *SEGNO*[119]
		yes	*RaFoSA*,[120] SACF,[121] *ASSP*,[122] *Method of Kneller and Hinsen*,[123] *PCASSO*,[124] *SST*,[125] *SABA*,[126] *PMML*,[127] *PROSIGN*,[128] *PALSSE*,[129] *Method of Taylor et al.*,[130] Method of Zhang and Skolnick,[71] *VoTAP*,[131] STICK,[132] *XTLSSTR*,[133] *P-SEA*,[134] *GETSSE*,[135] *P-CURVE*,[136] *DEFINE_STRUCTURE*[137]
Consensus Methods			*SKSP*,[138] TCM[139]

1.2.2 The Comparison of the Spatial Arrangement of Secondary Structure Elements

During the last decades, protein structure knowledge was largely exploited for modern drug discovery. Structure comparison methods enable the elucidation of protein relationships and the investigation of relationships on the level of domains, folds, and super-secondary structures.[140] A well-known example of a functional super-secondary structure motif is the Rossmann fold which is found to commonly interact with dinucleotide molecules (e.g., NADH, NADPH, or FAD).[141] Fold comparisons enable researchers to infer the function of a protein from its structural similarity and to track the enzymatic activities and regulatory mechanisms of yet uncharacterised proteins. The databases CATH,[142] Pfam,[143] SCOP,[144] and ECOD[145] facilitate the search for the fold of novel protein structures or domains. Intriguingly, also SSEs located at protein binding sites and their fold-independent 3D arrangements (ligand-sensing core[68]) seem to have an impact on protein function. It was previously shown that proteins sharing a common secondary structure organisation in the proximity of the ligand-binding site bind to similar molecular scaffolds.[68,146] This led to the development of the Protein Structure Similarity Cluster (PSSC)[68] workflow. In this workflow, the protein of interest is compared to the PDB using the Dali/FSSP server.[147] Pharmaceutically relevant protein families with low sequence similarity to the target are extracted from the ranked hit lists. After visual inspection, the catalytic cores of the target and the most similar proteins are aligned. Promising similarities of such subfolds might guide the way to new ideas for the identification of already validated ligands for the new target.

The approach was applied multiple times and the two most recent applications are shortly introduced here.

The first application was the search for a novel ligand and inhibitor of acyl-protein thioesterase 1 (APT1), an enzyme which hydrolyses fatty acids from S-acylated Cys residues with low lysophospholipase activity. A common ligand-sensing core of APT1 (PDB-ID 1fj2) with gastric lipase (PDB-ID 1k8q) was identified. Both structures and the corresponding cores show sequence similarities below 25%, but an overall high structural similarity. As lipase inhibitors are often characterised by a β-lactone core, β-lactone core-containing compounds were synthesised and tested for the inhibition of APT1. In consequence, two Palmostatins were identified as competitive inhibitors of APT1.[148]

Another interesting hint towards the applicability of the PSSC approach was the discovery of a potent nanomolar inhibitor of lysine-specific demethylase 1 (LSD1). The enzyme catalyses the post-translational modification of histones and regulates the androgen-dependent expression of genes involved in prostate cancer. Based on the finding that LSD1 (PDB-ID 2ejr) and the monoamine oxidases MAO-A (PDB-ID 2bxr) and MAO-B (PDB-ID 1gos) are structurally related by a similar arrangement of SSEs, γ-pyrones that were initially characterised as reversible inhibitors of monoamine oxidases were investigated for the inhibition of LSD1. The resulting hit, namoline (3-chloro-6-nitro-2-(trifluoromethyl)-4*H*-chromen-4-one), blocks the activity of the enzyme. It also reduces cell proliferation in a xenograft tumour model highlighting the potential impact of this compound on the attenuation of prostate cancer growth.[149]

Although the sequence similarity between the two protein pairs is low, both case studies are examples of protein domains with significant overall similarities which are already captured in databases (LSD1 and MAO-A/MAO-B: Pfam family “Amino_oxidase” - Flavin-containing amine oxidoreductase; APT1 and gastric triacylglycerol lipase: CATH-ID 3.40.50.1820 - “α/β hydrolase”). Therefore, the question arises whether a novel graph-based approach, which incorporates the complete PSSC workflow in one step, will enable the identification of common SSE arrangements in completely unrelated protein structures.

1.3 Trypanothione Synthetase

1.3.1 Neglected Tropical Diseases

Chagas disease (American trypanosomiasis), HAT (sleeping sickness), and leishmaniasis are caused by protozoan parasites of the order Trypanosomatida. These vector-borne diseases can be assigned to the large group of NTDs that affect more than one billion people in 149 countries with tropical and subtropical conditions.[1] Increasing knowledge of the lifecycles of the causative parasites and their molecular mechanisms of cell invasion has led to promising strategies for the development of safe drugs.[150]

HAT and Chagas disease are caused by two subspecies of *Trypanosoma brucei brucei* (*T. brucei gambensiense*, *T. brucei rhodesiense*) and *T. cruzi*, respectively. While the number of reported cases fell from 300,000 (1995) to 3,796 in 2014 for HAT, approximately eight million people are infected with *T. cruzi* and the parasite causes the death of more than 10,000 people per year. The transmission vector is the insect Glossina ("tsetse fly") and the triatomine bug ("kissing bug") for HAT and Chagas disease, respectively. A recent review focuses on the development of chemotherapeutic strategies for the treatment of HAT.[151] In the case of Chagas disease, the current treatment methods are restricted to the drugs nifurtimox and benznidazole which cause considerable side effects that often lead to an early stop of treatment. Additionally, drug resistance was observed in *T. cruzi*.[152]

Leishmaniasis is transmitted via various species of the female phlebotomine sand flies. The disease exists in three main forms: visceral (kala-azar, VL), cutaneous (CL), and mucocutaneous. Additionally, post-kala azar dermal leishmaniasis (PKDL) follows treated VL mainly in Sudan and India.[153] It was estimated that 50,000 to 90,000 new cases of VL occur per year – a disease that is fatal in over 95% of the cases without proper treatment. For CL, approximately 0.6 to 1 million people are affected by an infection per year. Current treatment methods use antimonial compounds, amphotericin A, pentamidine, and miltefosine. Their disadvantages are the high toxicity and the increasing multiple drug resistance. The disease-associated parasites which were of interest in the present work are given in Table 1.3.1 with the respective disease and its worldwide distribution.

Table 1.3.1. Leishmaniasis and its distribution. Three Leishmania species are given with the corresponding forms of leishmaniasis in mammals and their distribution.[154]

Leishmania sp. (Synonym)	Clinical Disease	Distribution
L. donovani (*L. archibaldi*)	VL, PKDL	Central Africa, South Asia, Middle East, India, China
L. infantum (*L. chagasi*)	VL, CL	North Africa, Mediterranean countries (Europe and North Africa), Southeast Europe, Middle East, Central Asia, North, Central and South America (Brazil, Venezuela, Bolivia, Mexico)
L. major	CL	Central and North Africa, Middle East, Central Asia

Crithidia fasciculata is a member of the order Trypanosomatida, but exclusively affects insect hosts.[155] Therefore, the organism is a suitable research organism for the safe investigation of intracellular parasites. Many inhibitor studies focused on the targets of this organism.

1.3.2 Trypanothione Synthetase as a Promising Drug Target

TryS is a unique enzyme of organisms of the genera Trypanosoma and Leishmania. It catalyses the biosynthesis of trypanothione ($T(SH)_2$, N^1,N^8-bis(glutathionyl)spermidine) – an essential redox metabolite whose function is comparable to that of the human glutathione (GSH). The physiological impact of $T(SH)_2$ for the protozoan parasites is diverse and includes antioxidant defence, DNA synthesis, detoxification, and metabolic regulation via disulphide reduction. These functions are mediated by enzymes such as tryparedoxin, glutathione peroxidase-like tryparedoxin peroxidases, or ascorbate-dependent peroxidase.[156]

It was shown that TryS is crucial for the survival of *L. infantum* and *T. brucei*.[157,158] $T(SH)_2$ is synthesised via two steps from the substrates GSH and spermidine or GSP, respectively. The ATP consumption results from the activation of GSH as illustrated in Figure 1.3.1. The second reaction step involves two possible substrates (N^1-glutathionylspermidine and N^8-glutathionylspermidine). It was shown that both substrates are accepted, but that there is a preference of N^8-glutathionylspermidine (GSP) over N^1-glutathionylspermidine in the case of TryS from *C. fasciculata*.[159] Furthermore, a substrate inhibition by GSH is observed to a varying extent for TryS from different organisms. The N-terminal domain of TryS catalyses the reverse reaction, i.e., the hydrolysis of $T(SH)_2$ to form spermidine and GSH via GSP.

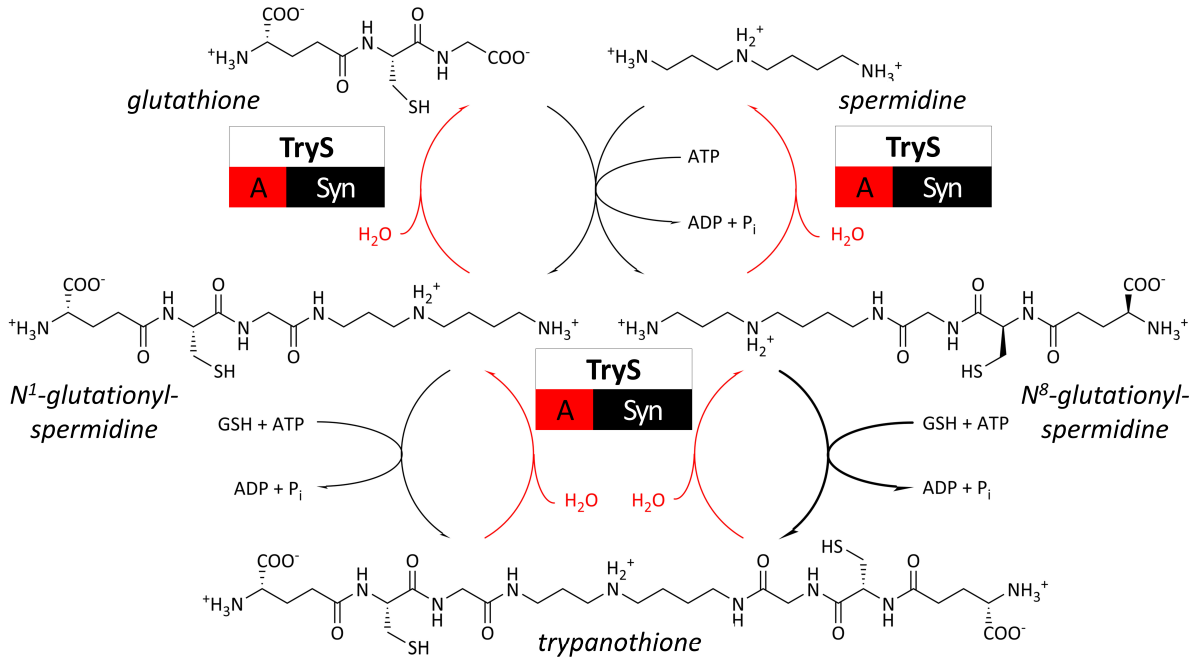


Figure 1.3.1. Trypanothione biosynthesis as catalysed by trypanothione synthetase. The black arrows indicate the catalysis by the C-terminal synthetase domain (Syn) and the red arrows represent the hydrolysis of $T(SH)_2$ and GSP catalysed by the amidase domain (A).

Although X-ray structures of the enzyme are published (PDB-IDs 2vob, 2vpm, 2vps),[160] the binding modes of the substrates and inhibitors have not been experimentally verified yet. The recombinantly expressed and purified enzyme from *L. major* consists of two domains: the N-terminal amidase domain and the C-terminal synthetase domain. The amidase domain is PROSITE[161]-annotated as a peptidase belonging to the family C51 which is characterised by the catalytic Cys and the neighbouring His and Asp residues. The synthetase domain shows the typical ATP-grasp fold.

Comparative modelling and molecular docking helped to gain insights into the binding mode of the substrates.[162] Figure 1.3.2 presents TryS from *L. major* in the presence of the substrates of the second reaction step (the predicted binding modes are shown). TryS and glutathionyl spermidine synthetase (GspS) from *Escherichia coli* (*Ec*GspS) are functionally related enzymes with sequence identities ranging from 30% to 33%. In contrast to monomeric recombinant TryS,[163–165] GspS crystallizes in the form of a dimer.[166] Figure 1.3.2 illustrates the structural similarity, especially with respect to the synthetase domains. This similarity enabled a modelling of the ATP accommodating cavity which is partly unresolved in known crystal structures due to its high flexibility. The final model of the enzyme shows the triangular arrangement of the three substrates around the reaction centre.

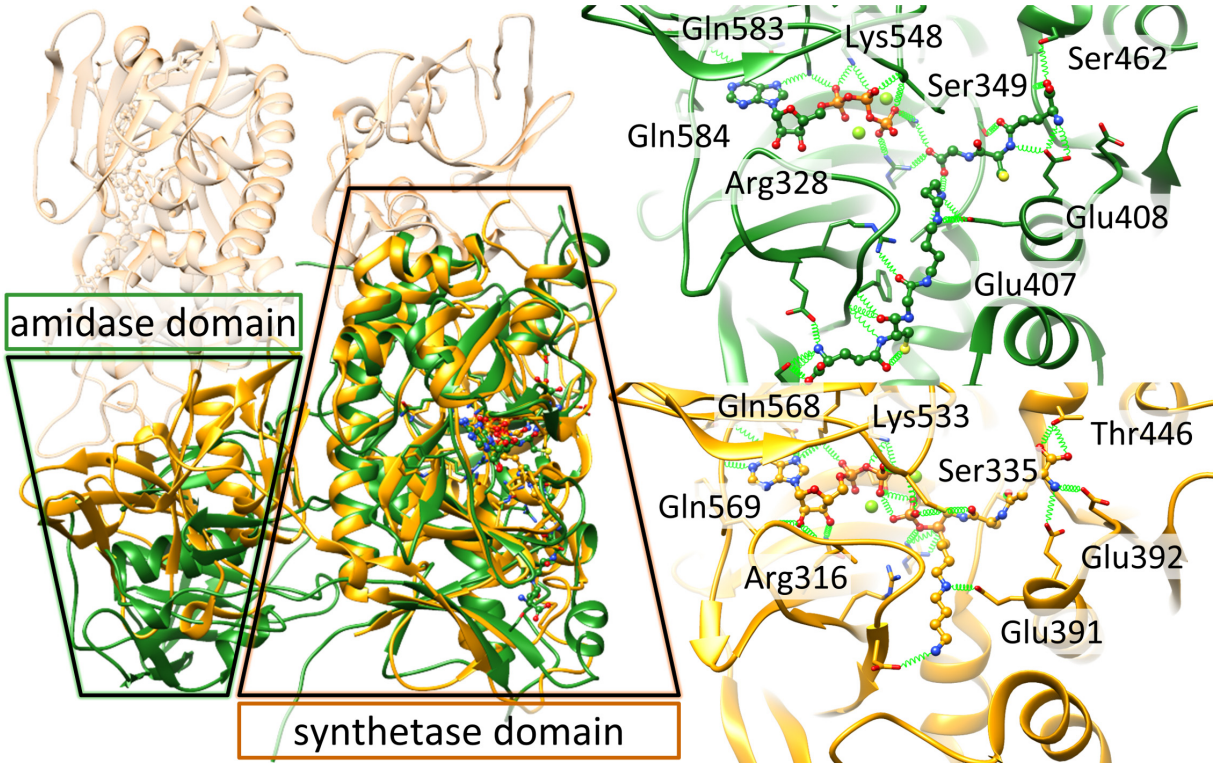


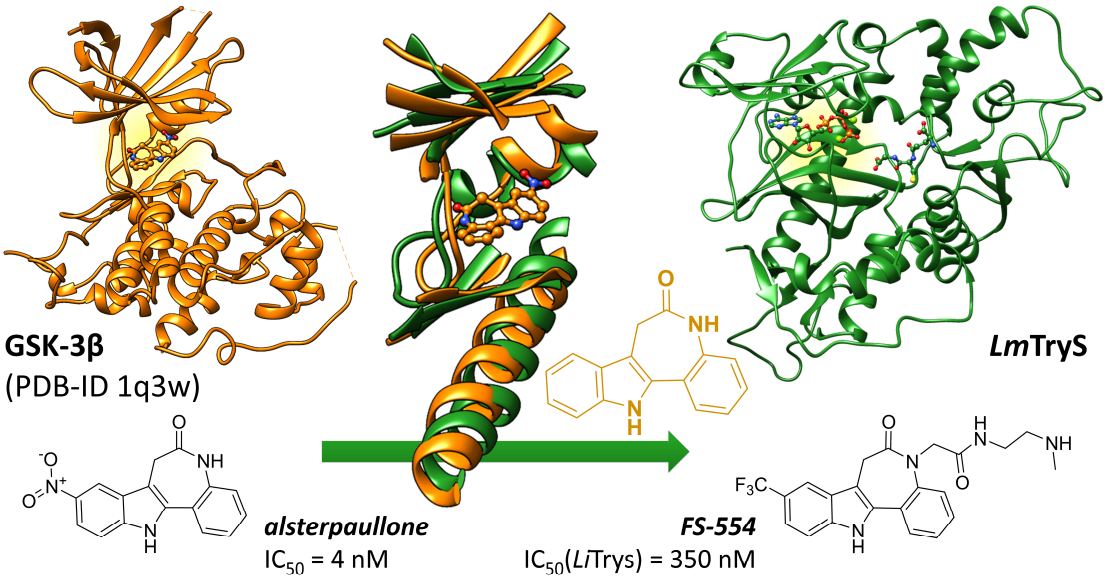
Figure 1.3.2. Comparison between the structures and active sites of *Ec*GspS and *Lm*TryS. On the left, the complete structures of the dimeric *Ec*GspS (orange, PDB-ID 2ioa) and the monomeric recombinant TryS (green, structure model based on the structure with the PDB-ID 2vps) from *L. major* are depicted. The alignment was generated for the synthetase domains of both structures. The amidase domain fold of both proteins is similar, but its relative arrangement with respect to the synthetase domain is different. Common residues of both enzymes, which interact with the substrates, are shown on the right. Green springs highlight hydrogen bond interactions. The substrates in the TryS active site are ATP, GSH, and GSP. *Ec*GspS is shown in complex with ADP and a GspS inhibitor ((2*R*)-2-amino-5-[[(2*R*)-1-[[4-(4-aminobutylamino)butylphosphonooxy-phosphoryl]methylamino]-1-oxo-propan-2-yl]amino]-5-oxo-pentanoic acid, ligand-ID: GGA). This inhibitor mimics GSH and spermidine. In contrast to GSH, the D-Glu and L-Ala are situated at the positions of L-Glu and L-Cys in GSH.

1.3.3 Trypanothione Synthetase and Protein Kinases

ATP grasp enzymes and protein kinases bind ATP and use the γ-phosphate either for the activation of reaction substrates in the case of ATP grasp enzymes[167] or the phosphorylation of substrates (in the case of kinases). Despite the distinct nature of both enzyme classes they share a common arrangement of SSEs in the proximity of the bound ATP.

This became evident after the analysis of the results of a high-throughput screening campaign for inhibitors of TryS from *C. fasciculata*. In the aftermath, it was shown that the compound MOL2008 (an N^5-substituted alsterpaullone, see Figure A4.1 for the molecular structures of the TryS inhibitors discussed herein) inhibits TryS from the organisms *C. fasciculata* and *L. infantum* in the nanomolar range. However, it shows only 40% to 60% inhibition of TryS from *T. brucei* and *T. cruzi* at a compound concentration of 30 µM.[168] A similar trend is observed for the paullone-based inhibitor FS-554. Nevertheless, the finding points towards a possible connection between both enzyme classes. This relationship might enable the identification of novel TryS inhibitors by the application of *in silico* VS tools to kinase inhibitor libraries. This similarity is in accordance with the concept of ligand-sensing cores. Figure 1.3.3 illustrates the SSE-based similarities of the enzymes TryS and glycogen synthase kinase-3β (GSK-3β).

The use of protein kinase inhibitors in repurposing studies for TryS might be a challenging issue due to the potential off-target activity towards human kinases. For the paullone scaffold, depicted in Figure 1.3.3, the substitution at the nitrogen atom engaged in a hinge hydrogen bond interaction in protein kinases leads to decreased activities on human kinases.[169] Similar results were shown for the compound series of 4-azapaullones tested for TryS inhibition.[170]




```
GSK-3β    58 DTKV  61 GSK-3β     70 VYQA    73 GSK-3β  82 VAIKKV  87
LmTryS   562 VVHA 565 LmTryS    555 IITS   558 LmTryS 540 GYAKKP 545

GSK-3β    98 LQIMRKL 104 GSK-3β  107 CNIVR 111  GSK-3β 128 ANLVLD 133
LmTryS   509 KAILPMI 515 LmTryS  523 PAILK 527  LmTryS 575 NMIYQQ 580

GSK-3β   154 PVIYVKLYMY QLFRSLAYIH SFG 176 GSK-3β      186 NLLLDPDTAV LKLCDF 201
LmTryS   264 SNHEFHLRCV AYGTQLHAIF MEA 286 LmTryS      326 DFAFNNETGE VKCFEY 341
```

Figure 1.3.3. The structural relationship between TryS and protein kinase GSK-3β. The SSE-based comparison of TryS in complex with ATP and protein kinase GSK-3β is presented at the top. The activity of both enzymes is inhibited by derivatives of the paullone scaffold and their corresponding ATP binding sites are characterised by similarly arranged SSEs. The paullone scaffold-based compound FS-554 Inhibits TryS but does not show high activities in a kinase panel screen. The sequence alignment of the superposed SSEs shows distinct dissimilarities of the residues which are part of the ligand-sensing core (bottom).

Although both enzymes show a similar arrangement of SSEs and, consequently, a similar orientation of binding site residues, the physicochemical properties of the binding sites differ considerably. As compared to alternative binding site comparison approaches, the comparison based on SSEs completely neglects the nature of the side-chains in the proximity of the ligand. As the binding mode of the paullone scaffold-based inhibitors MOL2008 and FS-554 is yet unknown, it is difficult to assess the impact of this common SSE arrangement on the binding of similar scaffolds. Therefore, the applicability of the binding site comparison tools, which were introduced in the previous section, was additionally analysed in this study. The application of a consensus approach which includes various binding site comparison approaches might lead to a more complete picture of the similarities and differences between the ATP binding sites of TryS and experimentally solved protein kinase structures.

2 AIM OF THE THESIS

The aim of this PhD thesis was the analysis of binding site comparison methods and the assessment of their impact on the identification of novel TryS inhibitors in the context of a drug repurposing strategy.

Small molecule binding sites contain important information for structure-based rational design. Therefore, a major question of the present work was: “What means binding site similarity?” Physicochemical binding site similarity, but also shape-based similarity can lead to similar ligand binding. Moreover, the potential impact of common ligand-sensing cores, i.e., a similar spatial arrangement of SSEs in the proximity of bound ligands, as an applicable idea generator for medicinal chemistry projects was evaluated. Therefore, a novel method was developed in close cooperation with a computer scientist. This method enables the automated graph-based comparison of secondary structure arrangements in an appropriate period of time. This algorithm in hand, it was analysed whether the concept of ligand-sensing cores can be applied for *in silico* SBVS.

For the comparison of binding sites, ligand-defined protein binding sites can be used. Nevertheless, it is often of interest to identify novel binding sites, assess their druggability, and find similar binding sites that lead to novel ideas for the identification of potential modulators of protein activity. The multitude of methods available for the detection and automated comparison of ligand binding sites hampers the decision for a suitable approach. Consequently, the question arises whether there is a tool that fulfils all prerequisites to successfully identify and rank binding site similarities. This work addresses this problem and analyses the benefits and drawbacks of different approaches to rationalise the choice of promising binding site identification and comparison tools. Additionally, the applicability of a consensus approach, that exploits the special strengths of various comparison methods, was investigated.

TryS – an interesting target for the treatment of neglected diseases – was used as the test case for the application of the analysed and developed computational tools. The enzyme catalyses two reactions leading to the product $T(SH)_2$ – a unique redox equivalent for parasites of the species Trypanosoma and Leishmania. The enzyme was shown to be crucial for the survival of *Leishmania infantum* and *Trypanosoma brucei* – organisms which cause the neglected diseases leishmaniasis and human African trypanosomiasis. Due to known similarities between TryS and protein kinases, a drug repurposing workflow was developed. Comparative modelling, MD simulations, and the resulting representative structural models comprised the starting points of all further analyses.

Two different binding site comparison approaches were applied. An elaborate workflow, which included the use of structural representatives, a consensus binding site comparison, molecular docking, and a final visual inspection, was performed to propose novel potential hit compounds worth testing for the inhibition of TryS in an *in vitro* assay.

3 METHODS

3.1 Binding Site Prediction

3.1.1 Analysed Standalone Software for Binding Site Prediction

Table 3.1.1 gives an overview of the analysed algorithms. The algorithms were chosen based on four prerequisites that render them applicable to a large-scale analysis of all publicly available experimentally solved protein structures: (a) the software can be applied locally on an up-to-date Linux workstation using the command line (standalone tools, no web servers or graphical user interface (GUI) based tools), (b) the analysed method is published, and the underlying algorithms and the basic usage are explained in a comprehensive manner, (c) the output offers distinct cavities and corresponding descriptors or at least pocket volumes which enable a subsequent ranking, and (d) the tool is applicable to novel protein structures. The fourth prerequisite does not only provide the possibility to process confidential data but also helps to predict pockets for representative sets of protein structures derived from simulations or structural models resulting from comparative modelling. Despite these restrictions, MetaPocket 2.0 was additionally analysed, as this method incorporates six (eight in the original publication) different prediction methodologies and gave extraordinarily high hit rates on different benchmark sets.[171]

ConSurf,[16] PDBinder,[17] and DEPTH (version 1.0)[172] report potential binding site residues across the whole protein without providing distinct pockets and their respective scores for cavity ranking. Similarly, proACT2,[173] GaussianFinder,[174] 3V,[175] and HOLLOW[176] report cavity pseudo-atoms without assigning them to distinct pockets. Therefore, they were excluded for further analyses, but might be interesting for a single POI with unknown binding sites. CAVER Analyst[177] enables the prediction of tunnels and pockets in proteins via a GUI. This tool was not used as the proposed pockets could not be exported. The same holds true for other GUI-based applications such as LIBRA.[178] The automated structure processing is infeasible for such methods.

Additionally, several tools could not be properly installed or executed due to insufficient documentation (e.g., Cryptosite[18]). In some cases, the tools rely on external software that is no longer available (e.g., eFindSite[179]). The HotPatch[180] software was not applicable due to insufficient documentation. Several binding site prediction approaches were included in integrated development environments (IDE) and not suited for casual users (GIRAF[181]). The methods dPredGeo and dPredGB[182] led to uninterpretable results during the initial testing. Another drawback is the necessity of an elaborate protein preparation and parameterisation for the subsequent simulation steps. These tools were omitted to ensure that the compared methods are appropriately applied. All publicly available web servers (except for MetaPocket 2.0) for binding site prediction were not used in these analyses as they cannot be applied in an automated manner for large protein datasets. The tools Pocket-Finder[183] and Q-SiteFinder[183] were not publicly available via the World Wide Web. A list of the approaches that have been established to identify binding sites can be found in Table A1.1 in Appendix 1.

Table 3.1.1 also provides information on the output format of the predictions. Pockets can be reported based on putative ligand-binding residues, grids that fill the identified cavity, or a single set of x-, y-, and z-coordinates for a point that indicates the location of a predicted site.

Table 3.1.1. Overview of the applied binding site prediction tools and their output format. Pocket locations can be provided as cavity centres (C), cavity filling grids or points (G), binding site residues (R). G-LoSA, S-SITE, TM-SITE, COFACTOR, COACH, MolSite, and Binding Response even report the potential ligands (L) of the identified similar binding sites.

Method	C	G	R	Availability (URL)
Geometry-Based				
BetaVoid[19]			x	http://voronoi.hanyang.ac.kr/software.htm#BetaVoid
CAVE[184]	x		x	http://cpc.cs.qub.ac.uk/summaries/AEHC_v1_0.html
CAVE-CL[185]	x		x	http://cpc.cs.qub.ac.uk/summaries/AEHC_v2_0.html
CAVITATOR[186]			x	http://cssb.biology.gatech.edu/Cavitator/
CAVITY[187]	x	x	x	ftp://162.105.160.5/pub/incoming/yxyuan/Programs/Cavity/
CAVITY SEARCH[188]		x		https://sourceforge.net/projects/cavity-search/files/
CLIPPERS (Travel Depth)[189,190]			x	https://github.com/ryancoleman/traveldistance
DoGSite[191]		x	x	https://www.biosolveit.de/DoGSiteScorer/
dxTuber[192]		x		https://github.com/willhelm-mueller/dxTuber
EDTSurf[193]			x	https://zhanglab.ccmb.med.umich.edu/EDTSurf/
EPOS^{BP}[194]		x	x	http://gepard.bioinformatik.uni-saarland.de/software/epos-bp
fconv[195]			x	http://pc1664.pharmazie.uni-marburg.de/download/
fpocket[196]		x	x	*fpocket (version1.0):* https://sourceforge.net/projects/fpocket/files/fpocket-1.0/fpocket-src-1.0/ *fpocket (version2.0):* https://sourceforge.net/projects/fpocket/files/latest/download
GetCleft[197]		x	x	http://biophys.umontreal.ca/nrg/NRG/Resources.html
ghecom[198]		x	x	http://strcomp.protein.osaka-u.ac.jp/ghecom/
KVFinder[199]		x	x	https://sourceforge.net/projects/kvfinder/
LIGSITE^{CS}[12]	x			http://projects.biotec.tu-dresden.de/pocket/download.html
McVol[200]		x		http://www.bisb.uni-bayreuth.de/index.php?page=data/mcvol/mcvol
MOLE 2.0[201]			x	https://webchem.ncbr.muni.cz/Platform/App/Mole
MSPocket[202]	x	x		http://projects.biotec.tu-dresden.de/MSPocket/
NanoShaper[203]		x		https://www.electrostaticszone.eu/downloads
PASS[11]	x			http://www.ccl.net/cca/software/UNIX/pass/overview.shtml
PHECOM[204]		x	x	http://strcomp.protein.osaka-u.ac.jp/ghecom/
PocketAnalyzer^{PCA}[205]		x	x	https://sourceforge.net/projects/papca/
PocketDepth[206]		x	x	https://sourceforge.net/projects/pocketdepth/
PocketPicker[207]		x		http://gecco.org.chemie.uni-frankfurt.de/pocketpicker/download.html
PrinCCes[208]		x		https://github.com/CzirjakGabor/PrinCCes
SURFNET[209]	x			http://www.ebi.ac.uk/thornton-srv/software/SURFNET/
trj_cavity[210]		x		https://sourceforge.net/projects/trjcavity/
VisGrid[211]	x	x	x	http://kiharalab.org/VisGrid/
Void[212]			x	upon request (P. Koehl)
VOIDOO[20]	x	x	x	http://xray.bmc.uu.se/usf/
VolSite[50]		x		http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html
Voronoia[213]	x		x	http://bioinformatics.charite.de/voronoia/

Table 3.1.1 *(continued)*. Overview of the applied tools and their output format. Pocket locations can be provided as cavity centres (C), cavity filling grids or points (G), binding site residues (R). G-LoSA, S-SITE, TM-SITE, COFACTOR, COACH, MolSite, and Binding Response even report the potential ligands (L) of the identified similar binding sites.

Method	C	G	R	Availability (URL)
Energy-Based				
AutoSite[214]		x		http://adfr.scripps.edu/AutoDockFR/downloads.html
Binding Response[14]		L		http://cadd.umaryland.edu/CADD_resources.shtml
MolSite[215]		L		http://presto.protein.osaka-u.ac.jp/myPresto4/
SiteHound[216,217]	x	x	x	http://scbx.mssm.edu/sitehound/sitehound-download/download.html
SiteMap[218,219]		x		https://www.schrodinger.com/sitemap *(commercial)*
Knowledge-Based				
COFACTOR[220]	x	L	x	https://zhanglab.ccmb.med.umich.edu/COFACTOR/
G-LoSA[221]		L		https://compbio.lehigh.edu/GLoSA/
LISE[222]	x			https://sourceforge.net/projects/msdock/files/
P2RANK[223]	x	x	x	https://github.com/rdk/p2rank
S-SITE[36]		L	x	https://zhanglab.ccmb.med.umich.edu/COACH/
TM-SITE[36]		L	x	https://zhanglab.ccmb.med.umich.edu/COACH/
Combined Approaches				
COACH[36]	x	L	x	https://zhanglab.ccmb.med.umich.edu/COACH/
ConCavity[34]		x		http://compbio.cs.princeton.edu/concavity/
DEPTH (version 2.0)[224]			x	http://cospi.iiserpune.ac.in/depth/htdocs/download.html
MetaPocket 2.0[171]	x	x	x	http://projects.biotec.tu-dresden.de/metapocket/

3.1.2 Dataset Preparation

Altogether, four datasets were prepared to assess the performance of the above listed binding site prediction methods. The extended PDBbind core dataset (including the PDBbind core complex structures[225] and their ligand-free counterparts), the set of cryptic site pairs (a subset of the cryptic site pairs used by Cimermancic and co-workers[18]), a set of comparative models of the *M. tuberculosis* proteome, and a representative dataset of allosteric binding sites (from the ASBench[226] web site).

The structures of the PDBbind core dataset (version 2013)[225] were visually inspected and aligned using PyMOL.[227] The structures in the dataset are characterised by a high model quality and span a diverse spectrum of protein families. For each representative protein, an additional apo structure was chosen to generate a dataset with occupied and unoccupied ligand binding sites. Although the search for apo structures focused on model quality, the dataset also includes some structures with a resolution above 2.5 Å and an R-factor above 0.25. For a detailed list see Table A1.2. This extension by ligand-free structures resulted in a set of four structures per member of the diverse protein dataset. Ions and small molecules from the crystallization buffer, water molecules, and cofactors were removed from the complex structures. The composite ligand consisting of all three ligands of the aligned protein-ligand complexes of each group was extracted and saved as a separate PDB file. All residues within 5 Å of these ligand atoms were defined as binding site residues for the subsequent performance evaluation.

The cryptic binding site pairs of the dataset of Cimermancic *et al.*[18] were downloaded from the PDB and aligned with the help of fconv.[195] Pairs of unbound and ligand-bound structures that led to inconclusive alignments were excluded from the analysis. Ligands from holo structures were also used as cavity reference for the ligand-free structures. Ligand atoms of holo structures that clashed with protein atoms of the corresponding apo structure were removed by means of a UCSF Chimera[228] script. This process led to the removal of all ligand atoms for apo structures whose binding sites differed notably from the ligand-bound ones. Therefore, these structure pairs were also excluded from the dataset. A complete list of structures and binding site-defining ligands of the final dataset can be found in Table A1.5. After these pre-processing steps, binding site-comprising residues were assigned using the extracted ligand atoms (all residues within a 5 Å radius).

To assess the impact of the inaccuracies of comparative models on the outcome of the binding site prediction, the homology models of the *M. tuberculosis* proteome were downloaded from the SWISS-MODEL repository.[229] These structures were filtered to obtain only models which were created with the help of templates with a maximum sequence similarity of 40%. The resulting models were further analysed with respect to bound ligands. Only structures in complex with ligands were considered. This workflow yielded 10 comparative models from the SWISS-MODEL repository and the 10 corresponding template structures which were downloaded from the PDB. Model and template structures were aligned. Template chains which were not used for comparative modelling were removed. The ligands of the templates were used as the reference for the binding site residue definition (in a 5 Å radius of the ligand atoms). Two ligands were included in the model for the template structure with the PDB-ID 4cvm leading to 22 binding sites in the dataset. The complete list of analysed structures, the corresponding ligands, and the model-IDs can be found in Table A1.9.

Finally, the performance of the methods for sites of allosteric regulation was assessed using the ASBench Core Diversity Set.[226] The binding sites of the 147 allosteric modulators were extracted based on a 5 Å radius of all ligand atoms. The protein chains were assigned after visual inspection to ensure that all chains constituting the binding sites are included in the analysis (see Table A1.11). Ligand-free structures were not used in this dataset.

3.1.3 Binding Site Identification - Necessary Pre-Processing Steps and Usage Details

The usage details for all methods are described below. Default parameters were applied whenever possible.

Additionally, the post-processing of the predicted binding sites included the assignment of binding site residues for tools which did not provide this information. In case of tools which report a single point as the pocket centre, residues in a 7, 8, 9, 10, 11, 12, and 13 Å radius were assigned as binding residues. The radius which led to the best hit rates for the extended PDBbind core set was finally chosen for evaluation purposes. Similarly, radii of 3, 4, 5, 6, and 7 Å were used for cavity grid-reporting tools.

Geometry-Based Methods

BetaVoid[19] BetaVoid was applied with the default settings. The properties of voids in the Lee-Richards model and the solvent accessible surface (SAS) area were calculated using a solvent radius of 1.4 Å. Voids were ranked by their volume.

CAVE[185] After the preparation of the input structure, the cavities were identified using the default parameters. The optimised version of the original CAVE algorithm[184] was used. The GPU-accelerated version CAVE-CL was applied in the same way and led to identical results.

CAVITATOR[186] The original source code restricts pocket prediction to proteins with a maximum of 320 residues. This parameter was changed to 2,000 to ensure the processing of all POIs (maximum number of residues: 1,016). The parameter ss/scalc was increased from 1.15 to 2.0 leading to the processing of proteins with a high radius of gyration. The residues of the structures were renumbered prior to pocket identification. Additionally, TER records separating two chains were removed.

CAVITY[187] The tool was used with the parameters to detect pockets in general (the additional parameter sets for shallow, complex, or very large binding sites were not evaluated). The output includes the pocket surface points, the grid points, and the pocket-lining residues. The latter two were used for the evaluation.

CAVITY SEARCH[188] The software was applied with the default settings. The resulting clustered cavities and crevices were ranked by their volume. The reported grids were used to define the binding site residues in a 3 Å radius.

CLIPPERS[190] First, TravelDepth was applied to generate TST and PHI files and to calculate the convex hull and the travel depth ("tstTravelDepth.py"). Surface pockets were mapped by using "nocav.tst"/ "nocav.phi" and "cav.tst"/ "cav.phi" files ("tstPocketMap.py"). The evaluation was based on the CAV files as their use led to the most promising results. The script "tm3pickerPdb.py" was applied to extract the identified pockets. The output included the identified cavities for the three different parameter sets which are given in Table 3.1.2. The pockets were ranked by their volume. The best results were obtained for the parameter set 2 (highest hit rates for the extended PDBbind core dataset).

Table 3.1.2. Parameter sets for the method CLIPPERS. All sets were used for benchmark purposes.

Parameter Set	Maximum Apolar Surface Area Fraction	Maximum Polar Surface Area Fraction	Minimum Volume [Å³]	Maximum Volume [Å³]	Minimum Apolar Surface Area [Å²]	Minimum Polar Surface Area [Å²]
set 0	0.8	0.5	300	800	150	100
set 1	0.9	0.7	200	1,000	100	50
set 2	0.9	0.9	200	1,000	50	50

DoGSite[191] The tool (version 2.0.0) was applied with the default parameters. The output files for the pocket grid ("_gps_P_") and the residue atoms ("_res_P_") were used for the evaluation. The options "-d" and "-l" were enabled to calculate the site descriptors and the druggability. The predicted sites were rescored with ALLO[230] to identify allosteric pockets.

dxTuber[192] The protein structures were prepared using UCSF Chimera.[228] The preparation steps included the addition of hydrogen atoms and the placement of the protein in a TIP3P water-filled box. Subsequently, VMD[231] was applied to create the DX potential maps as input for dxTuber. The command line version was used with cavity grouping and filling enabled. In addition to the 2D scanning method, the 1D scan methodology was tested. As the 2D method outperformed the 1D approach, results for the 2D approach are given. The adjustable parameter, which enables the filtering based on the number of voxel neighbours, had no impact on the results. The output file contains the grids of predicted pockets. A 4 Å radius was applied to assign binding site residues for the benchmark analyses.

EDTSurf[193] EDTSurf was used with the default parameters. Cavity files with the binding site-lining residues of the protein were used for evaluation purposes. The pockets were ranked by their volume.

EPOS^{BP}[194] The tool was applied with the standard settings.

fconv[195] Two types of pocket detection algorithms are provided in fconv. The first one is grid-based, while the second one exploits a Delaunay triangulation. The default options were retained for both methods.

fpocket[196] The default parameters were applied to detect binding sites with fpocket versions 1.0 and 2.0 (dev).

ghecom[198] The software was applied in the “p”ocket mode which corresponds to the extension of the prediction methodology introduced by Kawabata and Go,[204] but also in the “P”ocket mode, i.e., pockets are detected according to Masuya and Doi.[232] The results of both modes were comparable. Moreover, the detection mode for multiscale pockets (deep and shallow pockets) was applied which led to the worst results for the extended PDBbind core dataset. Consequently, the results of ghecom with mode “-p” were used for comparison purposes.

KVFinder[199] The default settings were used and pocket grids were extracted from the “output.pdb” files. The predicted sites were ranked by the pocket volume. The binding site residues were assigned according to the “results.txt” file.

LIGSITE^{CS}[12] The default parameters were used and the three pockets with the largest volume were analysed. The output is restricted to pocket centres. Based on the analysis of the hit rates using the residues within different radii of the site point of the centres, a 9 Å radius was chosen to extract the predicted binding site residues.

McVol[200] McVol predictions rely on PQR files which were generated using the software pdb2pqr[233] (version 2.1.0).[234] The AMBER force field was used for the generation of the PQR files. As this step failed for structures with missing atoms, MODELLER (version 9.18)[235] was additionally used to add missing atoms to these structures. McVol was applied with the default settings.

The usage of the resulting cavity files ("cav_i.pqr") led to worse results than the usage of the cleft files ("real_cleft_i.pqr"). The inclusion of both cavity types also led to less promising results. Therefore, the evaluation was performed for the cleft files which were ranked by the number of cavity grid points. As the output does not include binding site residues, they were defined as the residues within a 4 Å radius of the pocket grid.

MOLE 2.0[201] The software was applied using the standard settings (a probe radius of 3 Å, an interior radius of 1.25 Å, and a mesh density of 1.33 Å) for the cavity detection. The detection of tunnels was disabled.

MSPocket[202] MSPocket (version 1.1) was used with the default parameters. Pocket grids were assigned using the "vtx.pdb" files, while site-defining residues were extracted from the "atm.pdb" files.

NanoShaper[203] All protein atoms of the PDB files were extracted and the protein structures were saved as XYZR files. The parameterisation was performed with the tleap module of the AmberTools14 package[236] (ff12SB force field). NanoShaper was run with the default settings and the pocket detection was enabled (large probe radius of 3 Å, small probe radius of 1.4 Å). The detection of interior binding sites was disabled. Residues in a 4 Å radius of the grid were used for the evaluation.

PASS[11] PASS (version 2.0.36) was applied using the standard settings. As only cavity centres were reported, the residues within 9 Å of the centres were assigned as the cavity-defining residues.

PHECOM[204] The default parameters were applied and the final probe clusters were used for the evaluation.

PocketAnalyzer^{PCA}[205] The default parameters were used to predict potential binding sites. The type of calculation ("—mode") was set to "sep", i.e., the input files were processed separately. No averaging or principal component analysis (PCA), which are useful options when analysing MD frames of protein simulations, were performed.

PocketDepth[206] The default settings were used and the site ranking was realised using the five binding site descriptors size (number of grid points), number of polar atoms, number of non-polar atoms, number of surface atoms, and depth factor. The predicted pocket clusters which were deleted in the prediction run were removed from the corresponding ranking lists. In a post-processing step the rank-files, which are named "clustered_numberofclusters_ sizeranking_polarranking_nonpolarranking_surfaceatomcount_surfaceatomcountpervolume _depthfactor.rank", were analysed. Clusters (pockets) were ranked according to all criteria except for the surface atom count per volume which is not recommended by the developers.

PocketPicker[207] The Python-based script "PocketPicker_1.0.py" was executed with PyMOL[227] for all protein structures. All residues within 3 Å of the pocket grid were assigned as the binding site residues.

PrinCCes[208] The default settings were used and the output BLD files for the resulting triangles were generated for all pockets separately. All residues within a 4 Å radius of the cavity grid were assigned as the pocket-defining residues for evaluation purposes.

SURFNET[209] The software was executed with the standard settings. The pocket centres were used for evaluation purposes. A 9 Å radius led to the best results for the binding site residue assignment.

trj_cavity[210] The tool was applied with the default settings and the output grids as stored in the PDB file format were used for the performance assessment. All residues within a 4 Å radius were assigned as the pocket residues. The burial state of the cavities to be identified is an adjustable parameter. A state of 6 denotes sites surrounded by the protein, while the burial states of 1 to 5 can be observed for other cavities. The default setting is 5. Setting this parameter to 4 or 6 decreased the hit rates. Therefore, the default settings were retained.

VisGrid[211] The standard settings were applied to extract putative binding sites with VisGrid.

Void[212] The tool, which is based on the publication of UnionBall,[212] was kindly provided by P. Koehl at the UC Davis. The original source code was modified to process protein structures with missing backbone atoms. Additionally, the original source code was used for structures completed using MODELLER (version 9.18).[235] The latter procedure was chosen for the evaluation as it led to the highest hit rates. Structures were renumbered as the software is not able to sequentially process residues with insertion codes. The recommended setting of 1.4 Å for the computation of a protein's SAS was used. The vdW (van der Waals) radii of all atoms were chosen as defined in the OPLS[237] force field. The pocket grids and residues were obtained by the post-processing of the resulting PML and POC files, respectively.

VOIDOO[20] The script "cavity.csh" was executed to obtain VOIDOO predictions and the resulting LOG files were used to rank the pockets according to their calculated average volume.

VolSite[50] VolSite relies on pre-processed MOL2 files which were created for all structures using the CSD (Cambridge Structural Database) Python API 1.3.[238] The tool was applied using the default settings and residues within a 4 Å radius of the predicted binding site grids were used for evaluation purposes.

Voronoia[213] The command line version of the software was used. First, packing density VOL files were calculated (default settings). These files were subsequently used to calculate the cavity centres and assign the cavity-neighbouring atoms with Voronoia.

Energy-Based Methods

AutoSite[214] Hydrogen atoms were added to the dataset structures using the tool Reduce as integrated into AMBER2014.[236] Subsequently, the PDBQT files were generated for all protein structures with Open Babel.[239] The residues in a 4 Å radius of the resulting pocket grids were assigned as the binding site residues.

Binding Response[14] The protein structures were processed with the pdbcur tool of the CCP4 software package[240] to remove alternative conformations, ANISOU entries, residues with zero occupancies, and hydrogen atoms. Subsequently, UCSF Chimera[228] was used to assign partial charges to all protein atoms and add hydrogen atoms. Binding Response was applied with the default settings. A library of 1,000 ligands was used for the docking. Different pocket ranking methods were tested: (a) according to the binding sites' size, (b) according to the binding response *Bp* for 90% of all compounds with the most favourable interactions (average binding response), (c) the average binding energy for all ligands. The latter ranking measure led to the highest number of correctly identified sites for the extended PDBbind core set and was applied for all datasets. The residues within a 5 Å radius of the reported pocket grid were defined as the pocket-linking residues.

MolSite[215] Several C-Shell scripts which are provided by the MolSite developers enable the automated protein preparation and generation of pockets for the molecular docking of a ligand library with 200 small molecules which are provided in addition to the MolSite binaries. The binding site residues were defined by a 5 Å radius of the reported pocket grid points.

SiteHound[217] SiteHound was applied with two different probe types in separate runs. First, a carbon probe was used to derive the molecular interaction fields (MIFs). Second, a phosphate probe was used to identify potential binding sites. As the prediction of pockets for structures with missing atoms or residues is not possible with SiteHound, MODELLER (version 9.18)[235] was used to model the locations and coordinates of missing residues and atoms for these protein structures. Notably, it was ensured that no missing atoms were part of the true binding sites. SiteHound was subsequently applied to all prepared structures.

SiteMap[219] All protein structures were prepared with the Schrödinger tool Protein Preparation Wizard.[241] The resulting MAE files were subsequently used for the binding site predictions with SiteMap[242] using the default settings. The points that are summed to compute the volume of the binding sites were used for evaluation purposes. The residues in a 3 Å radius of the pocket were used to evaluate the residue-based performance.

Knowledge-Based Methods

G-LoSA[243] Prepared PDB files, which do only include the ATOM and HETATM, as well as TER records, were used as input files. Chemical features and surfaces were assigned with the help of Java-based scripts. Additionally, VisGrid[211] was applied to generate SASs and predict putative sites for the subsequent comparisons. Binding sites whose proteins have more than 30% sequence identity with the query structure were excluded from the analysis. The binding sites are ranked according to their similarity to known binding sites and the ligands of the known binding sites are applied as the pocket definition. Finally, the ligands of the most similar binding sites are compared to each other based on their location in the query protein. Ligands which overlap by at least 70% with ligands of higher scored sites were excluded to avoid any redundancy. The complete processing workflow was adopted from the initial publication showing the suitability of G-LoSA for binding site identification.[243]

LISE[222] The tool was evaluated in the absence and presence of PSSMs (position-specific scoring matrices) for the POIs. The method PSI-BLAST[244] was applied to generate PSSMs based on a non-redundant sequence database (GenBank CDS translations, PDB, SwissProt, PIR, PRF excluding environmental samples from whole genome sequencing (WGS) projects; 106,376,657 sequences in the 11/2016 version). The default parameters were used for the binding site prediction runs.

PRANK/P2RANK[41,223] P2RANK and PRANK enable the prediction of binding sites and the rescoring of binding sites identified by fpocket and ConCavity. The original publication showed significant prediction improvements for fpocket (version 1.0) and ConCavity^{LIGSITE}. Both, the prediction and rescoring performance, were tested and analysed. The pocket predictions obtained by applying fpocket and ConCavity were rescored using PRANK. The default parameters were applied for all analyses.

Combined Methods

COACH[36] The I-TASSER 5.0[245] software was used to perform binding site predictions with COACH. The EC (Enzyme Commission) and GO (gene ontology) flags were set to false. The "homoflag" flag was set to "benchmark" to ensure the exclusion of homologous templates for the similarity-based site predictions, i.e., templates with a sequence identity above 30% were not considered. The output includes the results of the predictions by COACH, COFACTOR, S-SITE, and TM-SITE.

ConCavity[34] The multiple sequence alignments (MSAs) were required for all processed PDB structures to evaluate the binding site prediction performance of ConCavity.[34] The residue conservation scores were calculated for the MSAs derived from the HSSP[246] (Homology-derived StructureS of Proteins) files which were downloaded from the public FTP server of the CMBI.[247] The HSSP files contain information for one chain if multiple chains with identical sequences are found in one PDB entry. The Perl-based script "Hssp2Fasta.pl"[248] was used to generate files in the format that is supported by the Python-based script "conservation_score.py". The script is based on an earlier publication[249] and is provided together with the ConCavity software.[250] The tool was applied in the absence and presence of conservation scores for all three reimplemented pocket prediction methodologies (LIGSITE, PocketFinder, and SURFNET). The pocket extraction methods "topN" and "search" were evaluated. The option "search" led to an enhancement of the hit rates for ConCavity^{LIGSITE} and ConCavity^{PocketFinder} for the extended PDBbind core set and was finally chosen for the evaluation of both tools. In contrast, the option "topN" was used to get optimal results for ConCavity^{SURFNET}. The method "blur" was applied for the residue mapping with all three approaches. For the definition of the binding site residues, a 3, 5, and 5 Å radius of the grid was used for ConCavity using the LIGSITE, PocketFinder, and SURFNET methodology, respectively. These radii were adjusted to 3, 4, and 5 Å for the three methods if the data for residue conservation were used as additional input.

DEPTH (version 2.0)[224] The second version of DEPTH was applied with the default parameters. The MSAs were derived with blastpgp (BLAST version 2.2.22)[244] for the UniRef90 database. An additional script was provided by the developers to group residues within a 9.8 Å radius. This leads to the clustering of the predicted residues belonging to distinct binding sites.

MetaPocket 2.0[171] The PDB files of the protein structures were uploaded to the web server and the top3 predicted pockets were requested. The predictions of POCASA[251] and Q-SiteFinder[183] failed in all cases. The SURFNET predictions failed for a small number of structures (PDB-IDs 1dp0, 1ps3, 3bub, 3d4z, 3eir, 3i3b, 3l4w, 3muz, 3vd4, 2zku for the extended PDBbind core dataset, PDB-IDs 2wqp, 4hat, and 4hb2 for the set of the cryptic site pairs).

3.1.4 Evaluation of the Prediction Performance

The accuracy and the prediction success were assessed from a residue-based perspective and from a ligand-based point of view. The comparative evaluation of the prediction performance of all tools was largely hampered by the different output formats for the predicted cavities (see Table 3.1.1). A cavity grid-based evaluation was impossible as some of the tools do not provide this kind of data. Other tools report the cavity-lining residues or only the coordinates of the pocket centre.

Therefore, two evaluation methods that do solely rely on a cavity centre were applied. The pocket centre can even be obtained if the cavity grids or residues are reported. In these cases, the geometric centre of the cavity was determined as the mean of all grid point or residue atom coordinates. The general idea behind both evaluation approaches for one showcase structure of the extended PDBbind core dataset is depicted in Figure 3.1.1. Both approaches estimate the relative accuracy of the prediction and give values between 0 (no overlap between the predicted and the known binding site) and 1 (perfect overlap between the predicted binding site and the known one). Different thresholds for both measures were used to identify correctly predicted sites and the hit rates for these predefined thresholds were calculated.

Residue-Centric Evaluation

After aligning the holo and apo structures and assigning the respective ligand atoms as described above, the residues within 5 Å of any ligand atom were defined as binding site residues. The relative residue overlap (*RRO*) between those residues (cavity-defining residues, *CDR*) and the predicted binding site residues (*PR*) was evaluated using Equation 1.1.

$$RRO = \frac{|CDR \cap PR|}{|CDR| + |PR| - |CDR \cap PR|} \tag{1.1}$$

The *RRO* of 0 means that none of the predicted binding site residues is part of the ligand-defined cavity, while the *RRO* of 1 denotes the perfect overlap of the cavity-defining and predicted residues.

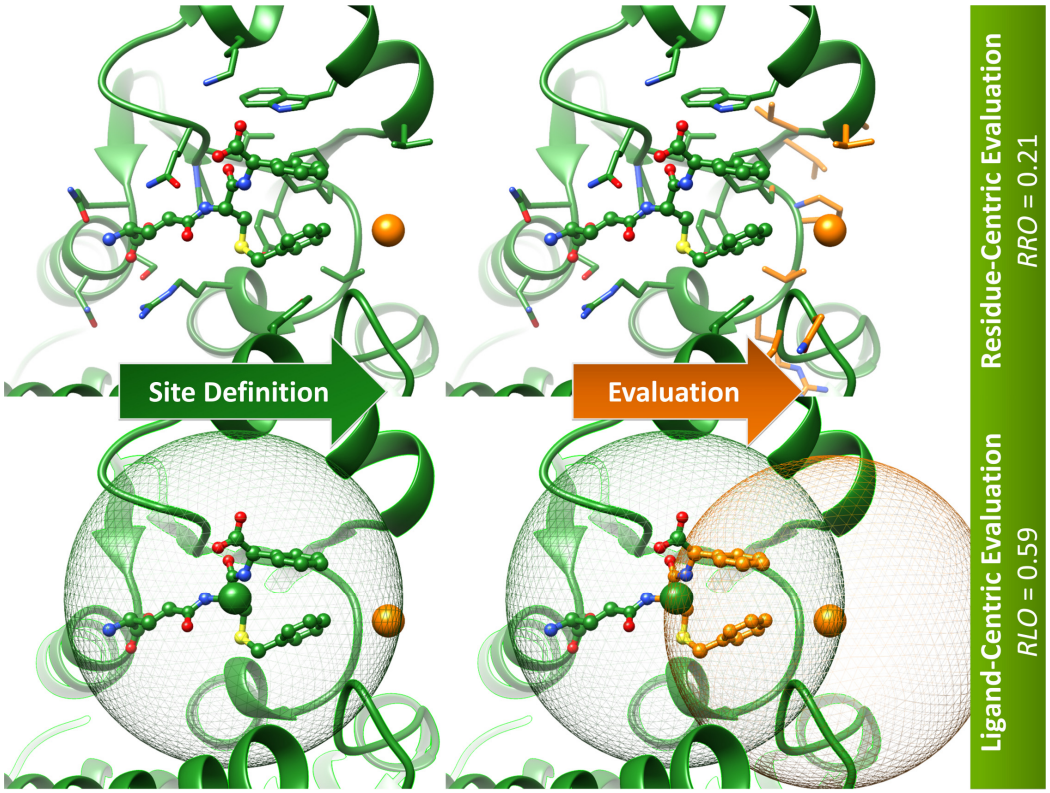


Figure 3.1.1. Residue-centric and ligand-centric methodology to assess the accuracy of binding site prediction methods. The small orange sphere corresponds to the detected site of the prediction method. Top: The residue-based analysis of the binding site prediction accuracy (*RRO*) is achieved by evaluating the overlap between the binding site residues (green, stick representation) and predicted binding site residues (orange, stick representation). Bottom: The precision of the prediction based on the ligand (*RLO*) is assessed by dividing the number of ligand atoms within the binding site-specific tolerance radius (indicated by the orange ligand atoms in ball-and-stick representation and the large orange sphere) from the predicted cavity centre (small orange spheres) by the number of all ligand atoms. The tolerance radius per pocket is determined based on the geometric mean of all pocket-defining ligand atoms (small green sphere). The tolerance is calculated from the largest distance between any ligand atom to this geometric mean plus 20% of this distance (large green sphere).

The predicted binding site residues for methods which do not report the corresponding binding site residues were extracted based on different distance thresholds. Residues within 3, 4, 5, 6, or 7 Å of any cavity grid point or within 7, 8, 9, 10, 11, 12, or 13 Å of the cavity-defining centre were extracted, respectively. The results for the different cut-offs were analysed and the distance threshold leading to the best results (highest hit rate for the top3 predicted binding sites for the PDBbind core set) was chosen for the evaluation.

A similar measure which is frequently used for the assessment of the prediction accuracy is the *MCC* (Matthews correlation coefficient)[252] which is calculated according to Equation 1.2. The numbers of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*) are captured in this performance metric. It considers the imbalance of a large number of non-binding residues and a small number of binding residues.[253] Furthermore, precision (Equation 1.3) and recall (Equation 1.4) are commonly analysed.

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{((TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN))}} \quad (1.2)$$

$$precision = \frac{TP}{TP + FP} \tag{1.3}$$

$$recall = \frac{TP}{TP + FN} \tag{1.4}$$

Both measures are predominantly suited for methods which report the binding and non-binding residues, while they are unsuitable for the evaluation of pocket grids and centres.

Ligand-Centric Evaluation

The prediction accuracy was also evaluated from a ligand-centric point of view. The maximum distance of all atoms of the respective ligand (ligand atoms, *LA*) to the geometric centre of all ligand atoms (ligand-specific radius, *LSR*) was defined. The tolerance radius (*TR*) was determined based on the equation $TR = LSR + 0.2 \cdot LSR$. Subsequently, the number of ligand atoms found within this ligand-specific radius of the predicted cavity centre (predicted ligand atoms, *PLA*) was calculated to determine the relative ligand overlap (*RLO*) by Equation 1.5. The value is defined between 0 (no ligand atoms included) and 1 (all ligand atoms included).

$$RLO = \frac{PLA}{LA} \tag{1.5}$$

The geometric mean of the predicted cavity-lining residues or the predicted cavity grid points were used for the methods which do not provide the respective cavity centre. If both residues and grids were reported, the grid was considered for the calculation of the geometric centre.

Evaluation of Prediction Robustness

The extended PDBbind core dataset includes four structures per target, enabling the assessment of the robustness of prediction, i.e., the fraction of correctly identified sites per target group. Methods which are highly sensitive towards conformational changes will be characterised by low robustness. Robust methods either identify all sites or none of them. The robustness per target group was set to 1 if none or all of the binding sites were correctly predicted, to 0.5 if one or three binding sites were correctly predicted, and to 0 if only two of the four binding sites were identified. The mean robustness and standard deviation for all groups per method were calculated.

Protein and Binding Site Alignments

The root-mean-square deviation (RMSD) values between the ligand-bound and unbound structures and their binding sites were calculated for all structures of the extended PDBbind core dataset and the set of the cryptic site pairs (see Tables A1.3 and A1.6 for the results). TM-align[71] was first applied for an alignment of the complete protein structures using the default settings. Then, the binding sites of the ligand-bound structures were defined as the residues in a 6 Å radius of the ligand (all three ligands in the case of the PDBbind core set). These binding sites were compared against the structures of the unbound proteins. Finally, ProBiS[61] was applied to compare the sites of the holo structures (all residues in a 5 Å radius of the ligand atoms were considered for the surface construction) against the apo protein sites.

Calculation of DoGSite Descriptors

The binding site descriptors and the druggability of all pockets of the extended PDBbind core dataset and the sets of cryptic and allosteric sites were calculated using DoGSite[254] (see Tables A1.4, A1.7, and A1.11). The pocket was defined by the geometric mean of the ligand atom coordinates. For the set of allosteric sites, the same procedure was followed, but all ligand atoms were used for the binding site definition.

3.1.5 Establishment of a Combined Approach to Improve Prediction Performance

The performance of all possible combinations of two to eight methods was calculated to combine complementary tools in a meta approach and assess its performance. This analysis was performed for the top1 and top3 ranked pockets to evaluate whether it is sufficient to use only the best scored pocket per method or the top3 scored pockets. Besides considering all methods, the analysis was also restricted to the tools which required less than half an hour per comparison on a single core basis (Intel Xeon CPU E5-2680 v2 @ 2.80GHz). Consequently, Binding Response, MolSite, LISE, G-LoSA, COACH (including COFACTOR, S-SITE, and TM-SITE), and DEPTH (version 2.0) were excluded for this second analysis.

The results of all tools for the binding sites of the extended PDBbind core[225] dataset and the cryptic site pairs[18] were analysed. Only the score of the method with the highest accuracy per pocket was considered per combination. Finally, the combinations were ranked according to the mean score for all pockets. Additionally, the combinations were compared based on the number of successful predictions (*RRO* or *RLO* ≥ 0.5) within the top1 and top3 ranked pockets.

Afterwards, the optimum combinations (combinations for the *RRO*- and the *RLO*-based evaluation) were analysed in more detail. The combinations of prediction tools which achieved the highest hit rates for the extended PDBbind core set were selected. The hit rates for these combinations for the dataset of the cryptic site pairs were subsequently evaluated. The combinations which led to the highest hit rate for this second set were used in a meta approach for the identification of binding sites. These two combinations were subsequently evaluated for all benchmark datasets in this study. The basic workflow of the meta approach is summarised in the following.

The top3 predicted pockets (pocket grids) of all tools were extracted. Then, duplicate pockets (distance between geometric pocket centres ≤ 5 Å) were removed starting with the pockets of the method which achieved the highest accuracy for the extended PDBbind core set. All remaining pockets were rescored according to the number of times they were found by the methods (FAVOUR score). Additionally, a rescoring procedure with PRANK[41] was performed. As this tool does only support the rescoring of fpocket- or ConCavity-derived pocket grid files, the pocket output files of all filtered predicted sites were transformed to the ConCavity pocket output format and the ConCavity-derived residue files were generated with ConCavity. These files were used as input for the PRANK rescoring procedure.

An overview of the tested meta approaches is given in Figure 3.1.2.

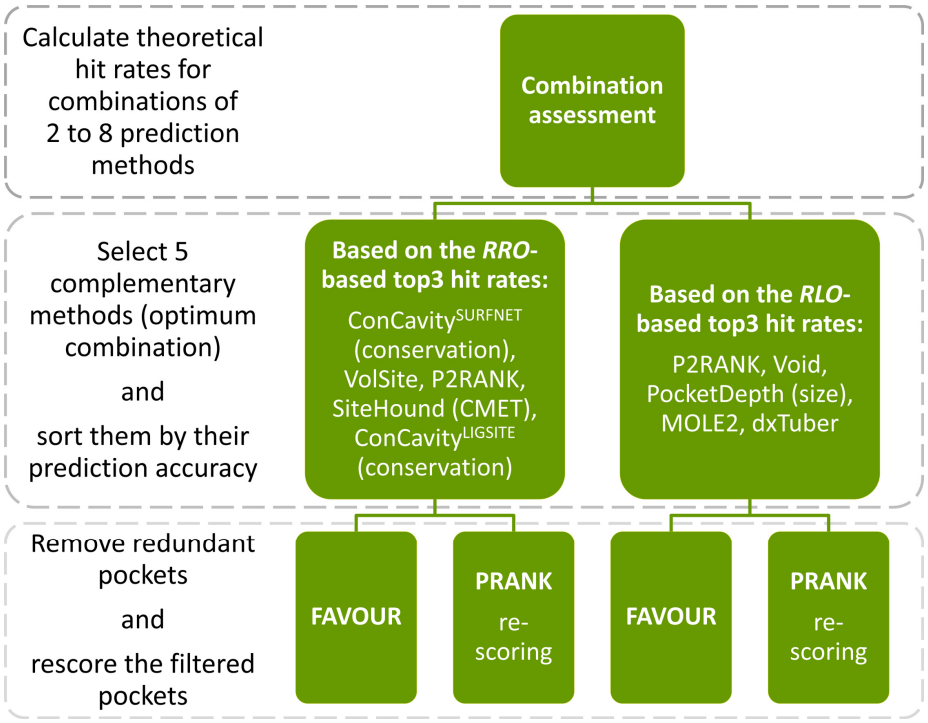


Figure 3.1.2. Overview of the tested method combinations and pocket rescoring schemes for the establishment of a meta approach to binding site prediction.

3.2 Binding Site Comparison

3.2.1 Dataset Preparation

The structures of all datasets were prepared equally to ensure a similar basis for all binding site comparison tools. First, modified residues were identified and the respective HETATM records in the PDB file were changed to ATOM records. The ligand which defined the binding site of interest was identified and renamed to LIG for the following processing steps. Other HETATM records in the PDB files were deleted to ensure the exclusion of buffer ions, cofactors, and prosthetic groups. The final steps were realised with the help of the pdbcur tool of the CCP4[240] software. In case of alternative locations of atoms, that one with the highest occupancy was retained, while the first one was retained in case of identical occupancy values. Finally, the ANISOU records were removed from the PDB files.

3.2.2 Dataset Characterisation

The report files for the structures of the datasets were downloaded from the PDB.[7] The resolution and R-factors of all structures were extracted. The mean values, standard deviations, minima, and maxima of these quality parameters were calculated for all datasets including X-ray structures. G-factors for the structures of all datasets were calculated using PROCHECK[255] and PROCHECK-NMR.[256] G-factors below -0.5 hint at unusual dihedral angles and/or covalent forces of the main-chain.

All residues in the protein structure groups of datasets 1, 1.2, and 2 (see below) were renumbered according to the sequence alignment calculated with MOE2013[257] using the default settings. Binding site residues were assigned in a 5 Å radius of the ligand atoms. The Cα atoms and all atoms of the binding site-defining corresponding residues of the structure pairs of one group were superimposed using the “match” command of UCSF Chimera.[228] The mean RMSD values, standard deviations, minima and maxima of all pairwise comparisons were calculated to characterise the binding site flexibility.

Binding site descriptors were calculated using DoGSite.[254] The binding site ligand was chosen as the reference ligand for the pocket definition. Apart from these changes, the default settings were used. The resulting pocket descriptors were analysed for each model of the NMR (nuclear magnetic resonance) ensembles of dataset 2, the sequence-culled set of the sc-PDB,[10] and the structures of dataset 7.

3.2.3 Datasets

Structures with Identical Sequences (Datasets 1/1.2)

The set of all single-chain entries of the sc-PDB[10] was sequence-culled using the PISCES[258] server with a sequence identity threshold of 25%. Only structures with a resolution of at least 3 Å and an R-factor of at least 0.3 were considered. The sequences of the resulting representative sc-PDB entries were compared to those of all single-chain protein structures stored in the PDB (08/2016) using USEARCH.[259] Structures from the representative set whose sequences were identical to those of at least nine other PDB-derived structures were retained.

This approach resulted in a dataset of 13 groups of structurally diverse proteins. The structures in the groups were aligned and the occupied binding sites were compared. This visual inspection led to the exclusion of one group represented by the sc-PDB entry with the PDB-ID 4l8u (chain A) as the location of the ligand-occupied binding sites differed for the structures with 100% sequence identity. An all-against-all comparison of all sites was performed to assess the performance of the analysed tools. Site pairs of proteins with identical sequences were regarded as "actives" (i.e., similar sites), while other matches were declared to be "inactives" (i.e., dissimilar sites).

Additionally, a reduced set of structures was generated which exclusively contains structures with identical sequences and similar ligands. To this end, the ECFP4 fingerprints[260] and the pairwise Tanimoto coefficients for all ligands per group were calculated in KNIME[261] and iteratively excluded the binding sites occupied with ligands with a Tanimoto coefficient below 0.6 to any other ligand in the group. Table A2.1 in Appendix 2 gives an overview of all dataset structures.

NMR Structures (Dataset 2)

A sequence culling using PISCES[258] was performed for all NMR structures in the sc-PDB.[10] Only ensembles which provide more than one conformer were considered. A sequence similarity threshold of 25% was set. Altogether 17 representative PDB entries containing different numbers of models were obtained.

An all-against-all comparison was performed and binding site pairs of the same model were regarded as similar sites. Although the NMR structures with the PDB-IDs 1diu (chain A) and 1yho (chain A) are homologous enzymes from two different organisms (dihydrofolate reductase, see Table A2.2 for an overview), the binding sites are dissimilar with respect to their residue composition. Therefore, binding site similarities between models of one enzyme should be ranked higher than similarities between models of both enzymes and matches between conformers of structures from both organisms were declared false positives.

Decoy Structures (Datasets 3/4)

One representative of each group of structures with identical sequences was chosen for the introduction of one, two, three, four, and five binding site residue substitutions. Randomly chosen surface-exposed binding site residues (random module in Python) were substituted by physicochemically different residues by means of the SwissSidechain[262] plugin of UCSF Chimera[228] (dataset 3). This plugin offers the possibility of choosing energetically favourable side-chain conformations of the newly introduced residues using the Dunbrack backbone-dependent rotamer library.[263] The ligand was retained in the binding site of interest. Finally, the resulting structures were protonated. Substitutions were designed in a way that the introduced residues were isosteric to the replaced residues, but showed different physicochemical properties (dataset 3). For the second dataset of decoy structures (dataset 4), the residues were randomly chosen for substitution and replaced by residues which differ by at least three in the number of hetero and carbon atoms from the original residue. Table 3.2.1 provides the substitution residues for the 20 standard amino acids.

Table 3.2.1. Decoy structure generation. Substitutions chosen to generate two datasets of rationally modified decoy structures for the evaluation of different binding site comparison algorithms are given.

Original Residue	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu
Substitution (Dataset 3)	*Gly*	*Ala*	*Leu*	*Arg*	*His*	*Ala*	*Phe*	*Thr*	*Glu*	*Asp*
Substitution (Dataset 4)	*Ile*	*Lys*	*Ala*	*Cys*	*Leu*	*Val*	*Ser*	*Ala*	*Ser*	*Ala*
Original Residue	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Substitution (Dataset 3)	*Gln*	*Leu*	*Ala*	*Met*	*Glu*	*Ala*	*Val*	*Thr*	*Phe*	*Phe*
Substitution (Dataset 4)	*Cys*	*Ala*	*Gly*	*Ala*	*Val*	*Lys*	*Gly*	*Gly*	*Phe*	*Ile*

The datasets were applied in two different ways that allowed for more sophisticated analyses. First, it was analysed whether the pairs of the original site structures from dataset 1 were assigned higher scores than pairs of the native sites and sites of the variants with one, two, three, four, or five substitutions. Next, the analysis was repeated by omitting all variants with less than five substitutions. The correlation between score and number of mutations was investigated by means of Spearman's rank correlation coefficient (see below for details).

The ROCS Dataset (Datasets 5/5.2)

The basic hypothesis that physicochemically similar ligands that bind in comparable conformations are indicative of related binding sites or a similar protein-ligand interaction pattern led to the development of dataset 5. It was set up to evaluate whether the binding site comparison tools can discriminate between binding site pairs binding to similar small molecules and those that bind to dissimilar ligands. The OpenEye tool ROCS (Rapid Overlay of Chemical Structures)[264] was applied to calculate the pairwise shape and physicochemical similarities for the ligands of the sites in the sc-PDB.[10] The complete structure library was downloaded (04/2017) and the ligand MOL2 files were used for an all-against-all comparison with ROCS. The results were filtered with respect to the TanimotoCombo score to obtain "active" (i.e., similar) and "inactive" (i.e., dissimilar) ligand pairs. "Active" pairs in this set show a TanimotoCombo of at least 1.4. "Inactive" pairs were defined as the binding site ligand pairs with a TanimotoCombo below 0.2. The corresponding protein structure pairs were compared using TM-align[71] to exclude proteins with high overall similarity. To this end, all chains involved in ligand binding as documented in the sc-PDB were used as input. Protein pairs with a TM-score above 0.3 were excluded. The resulting dataset comprises 15,339 "active" pairs of dissimilar proteins binding to similar ligands and 56,179 "inactive" pairs of unrelated proteins binding to dissimilar ligands (with respect to shape and physicochemical properties).

The Kahraman Dataset (Datasets 6/6.2)

The Kahraman dataset[265] consists of 100 protein-ligand complexes. They are listed in Table A2.3 with the corresponding ligands. The dataset includes inorganic phosphate ions as ligands. As phosphate is a small ligand undergoing a small number of interactions with the binding site residues, the performance of the tools for an all-against-all comparison of the dataset in the absence (dataset 6.2) and in the presence of the 20 phosphate binding sites was evaluated. Based on the idea of the Kahraman dataset, Hoffmann *et al.*[266] proposed a novel so-called "homogeneous dataset" with ligands of similar size. This dataset was not used in this study as it comprises not only ligands but also modified residues (PDB-based ligand-ID LLP) and unselectively binding ligands (ligand-IDs 1PE and BOG).

The Barelier Dataset (Datasets 7/7.2)

The Barelier dataset was designed using the results of the analyses of Barelier and co-workers[267] who analysed the relationships between binding sites of unrelated proteins binding to identical (or in one case similar) ligands. The structures were prepared by retaining all protein chains of the PDB entries. An overview of the structures and respective ligands can be found in Table A2.4. The structure with the PDB-ID 4odj comprises a single chain in the PDB entry. Nevertheless, the native protein can probably be regarded as a dimer. The ligand binds at the interface between two chains of the protein. Therefore, the dimer based on the symmetry copy was used to obtain the complete binding site. While the binding site pairs of class A were used as the similar site pairs, the remaining pairs (classes B and C) were used as the dissimilar site pairs. Finally, the cofactors of the proteins were included for the pairwise comparisons whenever possible (dataset 7.2).

The Dataset of Successful Applications (Dataset 8)

The most realistic dataset consists of binding site pairs which were previously characterised as being similar in the literature. The similarities as summarised elsewhere[48] were manually analysed and pairs or groups of similar binding sites were extracted (query sites). Altogether, 49 query sites were used (see Table A2.5). In addition, a sequence-culled dataset of the sc-PDB[10] structures was created with the help of the PISCES[258] web server leading to 1,103 structures at a sequence identity threshold of 25%. This representative dataset of sites was combined with the query structure (target sites) and compared to the query sites. The enrichment of the 115 similar cavity pairs in the similarity-ranked lists of the comparison methods was assessed.

3.2.4 Binding Site Comparison

The default parameters were used for all analysed tools unless otherwise stated.

If a method characterises binding site matches by more than one similarity or distance score, the scoring measure used for benchmarking was selected based on the early EF values for the datasets 1, 2, and 3 with descending priority (see below). However, all scores per method (Table 3.2.2) were calculated for the evaluation and the best performing measure per dataset is reported.

Extraction of Ligand Binding Sites

PocketMatch (version 2.0), SiteAlign (version 4.0), and TM-align (version 20170708) rely on pre-processed sites. The respective ligand-defined binding sites were extracted using a Python-based script which generates an output PDB file that contains all atoms within a given radius of the ligand. These excised pockets were used for the binding site comparisons.

A 5 Å radius was applied for the extraction of the input cavities for PocketMatch and SiteAlign. In contrast, a 10 Å radius was used for the retrieval of the binding sites for TM-align to ensure a sufficiently high number of residues which can be used for the initial sequence-based comparison.

For comparisons with FuzCav, VolSite/Shaper, TIFP, and Grim, additional MOL2 files of the structures were generated with the CSD Python API 1.3[238] and both, PDB and MOL2 files, were used as input. The name Shaper corresponds to the application of the method for MOL2 files, while the name Shaper(PDB) was used for the results if PDB files were used as input. This nomenclature was applied for all aforementioned methods.

Table 3.2.2. Summary of the scoring schemes provided by different comparison tools. An asterisk marks the scheme that was preferentially used for the evaluation of tools that offer multiple measures for the evaluation of the binding site similarity or distance.

Method	Scoring Schemes
Cavbase[64,65]	similarity score
FuzCav[73]	similarity score
Grim[268]	Grscore
IsoMIF [269]	tani*, taniM, taniMW, taniNormNodes, RMSD
KRIPO[74]	similarity score
PocketMatch[49]	$PMScore_{max}$*, $PMScore_{min}$
ProBiS[61]	z-score, SVA, RMSD, Alignment Score*
RAPMAD[270]	distance score
VolSite/Shaper[50]	Tanimoto (color, fit, combo*), RefTversky (color, fit, combo), FitTversky (color, fit, combo)
SiteAlign[77]	d1, d2, d3*, d4
SiteEngine[271]	LowResolutionScore, OverallSurfaceScore, DistanceScore, CurvatureScore*, TotalScore
SiteHopper[272]	ColorTanimoto, ShapeTanimoto, PatchScore*
SMAP [69]	Tanimoto, RawScore*
TIFP[268]	Tanimoto, Hamming, RefTversky, FitTversky, Dice, Soergel*
TM-align[71]	TM-score

Binding Site Comparison Software

Cavbase and RAPMAD[64,65,270] Both tools require XML-formatted pockets as input for the binding site comparison. The pockets were extracted using the CSD Python API 1.3[238] which was provided by the CCDC (Cambridge Crystallographic Data Centre). Proteins with residues with missing backbone atoms were not properly processed. Residues with missing backbone atoms were removed to obtain XML cavities if these residues were not part of the ligand binding site. The preparation failed for some proteins due to too long KEYWORD or JRNL records or unsupported space groups. Additionally, the cavity extraction for Cavbase- and RAPMAD-based comparisons is realised by the LIGSITE[273] cavity prediction approach. Therefore, the binding site of interest could not be found for some structures. Only the predicted pockets in the immediate proximity of the ligand atoms were compared. If more than one pocket for the ligand of interest was found, the pocket with the largest volume was used. The cavity extraction based on bound ligands is not possible. Additionally, very large cavities could not be processed using Cavbase. The use of RAPMAD is restricted to the similarity scoring of binding sites as the approach does not generate binding site alignments.

FuzCav[73] The binding site comparisons can be performed for the binding sites which were extracted with the pdbconv tool of IChem. The complete site (as used herein) or the residues within 4, 6, 8, 10, or 12 Å of the ligand can be used as input.

IsoMIF[269] Molecular interaction-based methods generally rely on bound ligands. A cut-off radius can be used to restrict the binding site dimensions depending on the ligand. Nevertheless, the use of the additional tool GetCleft[197] allows for the inclusion of predicted cavities for IsoMIF comparisons.

KRIPO[74] KRIPO fingerprint databases can be created by the users for individual, proprietary, and novel datasets. A ligand database is provided for all PDB entries. For proprietary or modified structures, the ligand database must be generated before preparing the fingerprints. The interaction fingerprints can be retrieved for ligand fragments and complete ligands with the command line tool “kripo”. Herein, the sites for the complete ligands were used. Subsequently, the KRIPO DB package (“kripodb”) was applied to perform all-against-all comparisons for the derived fingerprint databases. The resulting similarity matrix did not include the similarity scores for the identical site pairs. So, the identity scores were automatically set to 1.

PocketMatch[49] PocketMatch does not parse modified residues. The tool is suitable for similarity ranking, while the visualisation of the identified similarities is not possible. The PDB convention of 80 characters per line must be fulfilled for the PDB file-based input pockets and modified residues are not processed.

ProBiS[61] The binding sites are defined by the ligand structures. A distance threshold can be applied to adjust the cavity definition, but the default settings were applied herein, i.e., the residues in a 3 Å radius of the ligand atoms were used as pocket residues. ProBiS does not provide scores for all binding site pairs of interest. Cliques with poor scores or low z-scores are deleted and the sites are regarded as dissimilar. The “noprune” and “z-score” options offer the output of insignificant matches together with the significant ones. Both options were enabled leading to an increase in the number of site pairs with scores. However, the score of the site pairs which were not assigned a score was set to 0.

VolSite/Shaper[50] The use of Shaper relies on the cavity definitions by the VolSite method which predicts the druggability of binding sites and rejects non-druggable pockets. Therefore, some pockets cannot be processed by Shaper. Additionally, Shaper was used to compare the binding sites extracted by the pdbconv tool to allow for the more complete processing of the ligand-bound sites of the benchmark datasets. However, this is not recommended by the developers. The complete detected cavity and cavities that are defined within a 4, 6, 8, 10, or 12 Å radius of the ligand can be used for the comparison. The recommended 6 Å radius was applied for the analyses herein.

SiteAlign[77] Modified residues cannot be processed by SiteAlign. The binding sites have to be defined based on the residue names and numbers. Insertion codes are not supported necessitating the preliminary renumbering of the respective structures. The comparison of one query against a list of targets frequently fails with the error message “Segmentation Error”. Therefore, it is inevitable to perform the comparisons in a pairwise manner to avoid high failure rates.

SiteEngine[271] Alternate atom locations are not supported. A single conformation has to be kept in the PDB file. Moreover, the HETATM records for modified residues must be changed to ATOM records to ensure correct surface constructions. The tool does not process PDB files with more than 80 characters per line. Some large protein structures were not properly processed and had to be excluded. SiteEngine comparisons depend on ligand-defined cavities, i.e., its use is restricted to protein binding sites in complex with ligands. The distance cut-off for the binding site definition can be adjusted. However, the default setting of 4 Å was used for these analyses.

Sitehopper[272] SiteHopper initially creates binding site patches. This step failed for some protein-ligand complexes. Furthermore, the tool did not process residues with missing backbone atoms. These residues were therefore excluded for the comparisons after ensuring that they are not part of the binding site of interest. SiteHopper relies on ligand-defined binding sites, but the cut-off radius can be adjusted. For small ligands, an additional flag has to be set during protein-ligand splitting ("-min_atoms"). This flag was set to 0 to process the phosphate binding sites of dataset 6.2 and to 1 for the remaining datasets.

SMAP[69] SMAP (version 2.0) comparisons worked for nearly all structures after providing the PDB files in an appropriate manner. The tool often failed due to certain JRNL records which were therefore removed from all structures. The binding site definition can be modified by defining a distance threshold. The default settings as supplied with the executable were used.

TIFP and Grim[268] These tools of the IChem software package (version 5.2.6) are based on interaction fingerprints and graphs. Only ligand-occupied cavities can be compared. The tool pdbconv of the IChem toolbox must be applied to extract proper binding sites and interaction points. Problems arose if the ligand of interest was not in the predefined template files. Therefore, the original PDB-derived ligand names were retained for the comparisons with both methods. The input structure preparation also included a PDB to MOL2 file conversion with the CSD Python API 1.3[238] to obtain more reliable fingerprints. Additionally, the original ligand names from the PDB were used.

TM-align[71] The algorithm relies on a residue- and SSE-based sequence alignment. The binding site atoms within 10 Å of the bound ligands were used to ensure meaningful comparisons. As the tool relies on an initial sequence alignment of cavity residues, the small number of residues in the immediate neighbourhood of the ligand atoms prevents the alignment.

3.2.5 Performance Assessment

The ranking lists of the binding site comparison tools were used to investigate their performance with respect to the area under the ROC curve (AUC) of the receiver operating characteristics (ROC) curve and the enrichment of similar site pairs. The enrichment factor (EF) describes the enrichment of similar binding site pairs opposed to the number of similar pairs which are identified in a random screening (Equation 2.1).

$$EF_{x\%} = \frac{N(actives)_{x\%} / N_{x\%}}{N(actives)_{100\%} / N_{100\%}} \quad (2.1)$$

The EF for *x%* of the screened dataset is calculated from the number of true actives at this percentage ($N(actives)_{x\%}$), the number of all pairs at this percentage ($N_{x\%}$), the number of true active pairs in the complete dataset ($N(actives)_{100\%}$), and the number of all pairs in the complete dataset ($N_{100\%}$).

For the tools with more than one scoring scheme, the early EFs for all scores were analysed in more detail. The score that led to the highest early enrichment for the identical structures was considered. If the results for the scoring measures did not differ, the results for the dataset 2 were compared in a similar manner. Finally, the EFs for the dataset of ration decoy structures (dataset 3) were considered. The finally applied scoring schemes can be found in Table 3.2.2.

The ROC curves were plotted with the help of the KNIME[261] ROC Curve node to analyse the sensitivity (true positive rate) and the specificity (true negative rate) of the tools (Equations 2.2 and 2.3).

$$sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{2.2}$$

$$specificity = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{2.3}$$

The statistical analyses of the AUC differences were performed according to DeLong *et al.*[274] using pROC[275] as implemented in the software package R.[276]

The notched boxplots of similarity scores and distance measures for the similar and dissimilar pairs of dataset 1 were generated using the software package R.[276]

The Spearman's rank correlation coefficient (Spearman's Rho, r_S) was calculated for the decoy binding site datasets (datasets 3 and 4) according to Equation 2.4.

$$r_S = \frac{\mathrm{cov}(rank_X, rank_Y)}{\sigma_{rank_X} \cdot \sigma_{rank_Y}} \tag{2.4}$$

The raw scores (X_i) and the number of mutations (Y_i) were converted to ranks ($rank_X$, $rank_Y$). The covariance of the rank variables divided by the product of the standard deviations of both rank variables gives the correlation coefficient. In this study, the correlation between the binding site similarity score and the number of binding site residue substitutions was calculated. The general perception is that the higher the number of binding site substitutions, the lower the score. Thus, if the Spearman's Rho equals -1, a perfect correlation exists, i.e., the similarity score increases with decreasing numbers of substitutions.

3.2.6 Runtime Analyses

All-against-all comparisons with the structures of dataset 6 were performed on an Intel Xeon workstation (E5-2690 with 2.90GHz and 32 GB RAM) in a serial manner (single core). The pocket preparation for the comparisons with SiteAlign, PocketMatch, and TM-align was realised externally. For ProBiS, preparation and comparison are performed on the fly. The time for the post-processing of the results is not given. The runtime was assessed with the Bash “time” command (user time).

3.3 Secondary Structure Assignment

3.3.1 Clustering of Turns According to Their Conformation

Some of the subsequently described approaches for the classification of turns were inspired by a previous work which analysed the impact of turns on secondary structure assignment.[277,278] The earlier procedure as implemented in SHAFT[106] relies on the pre-processed Secbase[277] entries. In contrast, we parse the original PDB files, extract different turn categories, and prepare the structures independently and by means of an automated software which was implemented by Dipl.-Inf. Tobias Brinkjost. The turn assignment procedure supports modified residues with only minor exceptions (see List A3.1 in Appendix 3 for the list of residues which are not used for the turn and SSE assignment). Different optimisation steps were applied, and different parameters were tested for the new turn clustering procedure which is included in our SSE assignment method SCOT (Secondary structure Classification On Turns).

Datasets for Turn Clustering and the Analysis of SCOT-Derived Secondary Structure Elements

We designed two datasets for the classification of turns. All publicly available X-ray structures as stored in the PDB (03/20/2017) were filtered based on a resolution threshold of 2 Å and a maximum R-factor of 0.25. Subsequently, the PDB-IDs of the 17,978 filtered structures were submitted to the sequence-culling web server PISCES.[258] An identity threshold of 35% led to a representative dataset of 3,597 X-ray structures of high quality. These structures were used as the training set for the classification of turns in protein structures.

A second dataset for testing purposes was generated from the initial high-quality dataset after the exclusion of the training set structures. The remaining 14,379 structures were submitted to sequence culling with the help of the PISCES[258] web server. The sequence identity threshold was set to 35%. The resulting 2,481 structures served as the test set to analyse the outcome of the application of the novel turn assignment to an external dataset. The prediction accuracy of the turn clustering based on ESOMs (Emergent Self-Organizing Maps) was evaluated by visually inspecting the classified turns from the test set. Additionally, the dihedral angles, hydrogen bond energies, and Cα-Cα distances of turns of one turn type were statistically analysed (see below).

The original training set contained only six reverse 2 turns which prevented the ESOM training and analysis of the underlying U-Matrix for this turn category. A PISCES sequence culling with an identity threshold of 95% of the original high-quality dataset led to eight reverse 2 turns. Therefore, we adjusted the quality criteria (a maximum resolution of 3 Å and a maximum R-factor of 0.3) and created a new sequence-culled training set (with a sequence identity threshold of 65%). Nonetheless, the resulting 28,170 protein chains including only 53 reverse 2 turns were not sufficient for a reliable ESOM-based clustering. The lower limit of neurons for reliable ESOM training is 4,000.[279] Therefore, we used all protein X-ray structures of the PDB (11/10/2017) excluding the sequence-culled training set. This training set was used as the test set in the case of reverse 2 turns.

Placement of Backbone Hydrogen Atoms

The hydrogen atoms of the main-chain nitrogen atoms were added to all protein X-ray structures according to the publication of McDonald and Thornton.[280] The hydrogen atom is placed in the plane defined by the atoms constituting the peptide bond, but 4° closer to the Cα atom than to the carbonyl carbon atom of the respective residue. Pro and some modified or atypical residues have alkylated nitrogen atoms. These residues were excluded from this procedure after the visual inspection of all residues in the MODRES and SEQRES sections of all files in the PDB.

The List A3.2 in Appendix 3 contains the PDB names of compounds with atypical backbone structures which can be found as part of protein polymers. Their nitrogen atoms were ignored during the placement of backbone hydrogen atoms.

Hydrogen Bond Energies

Different formulae at different levels of accuracy exist to evaluate the strengths of hydrogen bonds. Herein, five empirical approaches were analysed to finally decide on one method which provides optimum results with respect to the hydrogen bond geometry and the corresponding energy.

Kabsch and Sander[104] proposed the following model equation to define the energy of hydrogen bonds E_{hb} in kcal/mol:

$$E_{hb} = q_1 \cdot q_2 \cdot \left(\frac{1}{d_{ON}} + \frac{1}{d_{CH}} + \frac{1}{d_{OH}} + \frac{1}{d_{CN}} \right) \cdot f \qquad (3.1)$$

The terms d_{ON}, d_{CH}, d_{OH}, and d_{CN} describe the distances between the backbone oxygen atom and the backbone nitrogen atom, the backbone carbon atom and the backbone hydrogen atom, the backbone oxygen atom and the backbone hydrogen atom, and the backbone carbonyl carbon atom and the backbone nitrogen atom in Å, respectively. The constant f takes a value of 332 and q_1 and q_2 take the values $0.673 \cdot 10^{-19}$ C and $0.32 \cdot 10^{-19}$ C. In the corresponding publication, it was stated that a stable hydrogen bond has a binding energy of approximately -3 kcal/mol. A cut-off of -0.5 kcal/mol was introduced to exclude weak electrostatic interactions.

An extension of this definition was proposed by Zacharias.[281] Equation 3.2 was introduced and applied for the analysis of backbone hydrogen bonds during the development of the PSSC (Protein Secondary Structure Classification)[112] approach.

$$E_{hb} = \left(\frac{q_O q_H}{d_{OH}} + \frac{q_C q_H}{d_{CH}} + \frac{q_N q_C}{d_{NC}} + \frac{q_O q_N}{d_{ON}} + \frac{q_C q_{C\alpha}}{d_{CC\alpha}} + \frac{q_O q_{C\alpha}}{d_{OC\alpha}} \right) \cdot f \qquad (3.2)$$

The partial charges defined in the CHARMM[282] force field were used (q_O = -0.51 C, q_N = -0.47 C, q_C = 0.51 C, $q_{C\alpha}$ = 0.16 C, q_H = 0.31 C). A hydrogen bond energy threshold of -0.75 kcal/mol was applied to assign hydrogen bonds.

The definition of the hydrogen bond potential used by Boobbyer and co-workers incorporates the angle dependence of hydrogen bond interaction strengths.[283] It was originally designed for use in the GRID energy function.[284] Nevertheless, Frishman and Argos used an adjusted version to describe backbone hydrogen bonds for the development of the SSAM STRIDE.[105] It includes an additional angular function that was modelled and fitted to experimental data. The hydrogen bond potential is defined by

$$E_{hb} = E_r \cdot E_t \cdot E_p \tag{3.3}$$

E_r is a function of the distance, while E_t and E_p reflect the spatial orientation of hydrogen bond donor and acceptor atoms as illustrated by Equations 3.4 to 3.6.

$$E_r = \frac{-3E_m \cdot r_m^8}{r_{ON}^8} - \frac{-4E_m \cdot r_m^6}{r_{ON}^6} \tag{3.4}$$

$$E_t = \begin{cases} (0.9 + 0.1\sin 2t_i)\cos t_0 & ; \quad 0 < t_i < 90° \\ K_1 \cdot (K_2 - \cos^2 t_i)^3 \cos t_0 & ; \quad 90° < t_i < 110° \\ 0 & ; \quad t_i > 110° \end{cases} \tag{3.5}$$

$$E_p = \cos^2 t \tag{3.6}$$

E_m and r_m are set according to the optimum hydrogen bond energy (-2.8 kcal/mol for N-O hydrogen bonds) and donor-acceptor distance (3 Å for N-O hydrogen bond interactions), respectively. The angle t_0 is defined by the hydrogen atom, the acceptor atom, and the intersection between the plane of the peptide bond and its perpendicular to the hydrogen atom. The angle t_i is defined between the C-O axis and the straight line between the oxygen atom and this intersection. K_1 and K_2 are part of the smoothing function and are defined by $0.9/\cos^6 110°$ and $\cos^2 110°$, respectively. The equation for the calculation of E_t corresponds to that for sp^2 hybridised carbonyl and carboxyl oxygen atoms as hydrogen bond acceptor, while E_p is used for an NH-type hydrogen bond donor. Correspondingly, the angle t is defined between the N-H axis and the straight line between the hydrogen and the acceptor atom. In the original publication, E_p is set to 1 assuming that the probe orientation in the GRID approach will correspond to the strongest hydrogen bond.

The hydrogen bond potential of the DREIDING[285] force field with minor adjustments of the angle terms was used by Dahiyat *et al.* to analyse α-helices.[286] Equation 3.7 describes the DREIDING hydrogen bond term. Equations 3.8 and 3.9 describe the hydrogen bond energy potential used by Dahiyat and colleagues.

$$E_{hb} = D_{hb} \cdot \left[5 \cdot \left(\frac{R_{hb}}{r_{ON}} \right)^{12} - 6 \cdot \left(\frac{R_{hb}}{r_{ON}} \right)^{10} \right] \cdot \cos^4 \theta_{OHN} \tag{3.7}$$

D_{hb} and R_{hb} were defined with 9.5 kcal/mol and 2.75 Å, respectively.

$$E_{hb} = D_{hb} \cdot \left[5 \cdot \left(\frac{R_{hb}}{r_{ON}} \right)^{12} - 6 \cdot \left(\frac{R_{hb}}{r_{ON}} \right)^{10} \right] \cdot F(\theta_{OHN}, \theta_{COH}, \varphi) \tag{3.8}$$

$$F(\theta_{OHN}, \theta_{COH}, \varphi) = \begin{cases} \cos^2 \theta_{OHN} \cos^2(\theta_{COH} - 109.5°); sp^3\ donor - sp^3\ acceptor \\ \cos^2 \theta_{OHN} \cos^2 \theta_{COH} \quad ; sp^3\ donor - sp^2\ acceptor \\ \cos^4 \theta_{OHN} \quad ; sp^2\ donor - sp^3\ acceptor \\ \cos^2 \theta_{OHN} \cos^2[\max(\theta_{COH}, \varphi)]; sp^2\ donor - sp^2\ acceptor \end{cases} \tag{3.9}$$

The angle φ is defined by the angle between the normals to the planes that are spanned by the atoms attached to the sp^2 centres. For this hydrogen bond potential, D_{hb} = 8 kcal/mol and R_{hb} = 2.8 Å were used. Additionally, hydrogen bond potentials are only evaluated if the donor-acceptor distances are between 2.5 and 3.3 Å and the donor hydrogen-acceptor angle is at least 130°.

The last evaluated function was derived as a semi-empirical force field equation to improve crystal structure modelling by Fabiola and co-workers.[287] This hydrogen bond potential includes a directional component that ensures the consideration of two optimum H-O-C angles. Equation 3.10 gives the hydrogen bond potential and Equation 3.11 describes the associated switching function as adopted from Brünger[288] which smoothly decreases the potential to 0 between the cut-off distances r_{ON}^{on} = 3.4 Å and r_{ON}^{off} = 3.5 Å.

$$E_{hb} = \varepsilon \cdot \left[\left(\frac{K}{r_{ON}} \right)^6 - \left(\frac{K}{r_{ON}} \right)^4 \right] \cdot \cos^4\left(\theta_{CON} - \theta_{CON}^0 \right) \cdot SW\left(r_{ON}, r_{ON}^{on}, r_{ON}^{off} \right) \tag{3.10}$$

$$SW\left(r_{ON}, r_{ON}^{on}, r_{ON}^{off} \right) = \begin{cases} 0 & ; \quad r_{ON} > r_{ON}^{off} \\ \dfrac{\left(r_{ON}^{off\,2} - r_{ON}^{2} \right)^2 \cdot \left(r_{ON}^{off\,2} - r_{ON}^{2} - 3 \cdot \left(r_{ON}^{on\,2} - r_{ON}^{2} \right) \right)}{\left(r_{ON}^{off\,2} - r_{ON}^{on\,2} \right)^8} & ; \ r_{ON}^{on} < r_{ON} < r_{ON}^{off} \\ 1 & ; \quad r_{ON} < r_{ON}^{on} \end{cases} \tag{3.11}$$

The constant K is calculated as $r_{ON}^{0} \cdot \sqrt{2}/3$ with r_{ON}^{0} as the optimum acceptor-donor distance of 2.9 Å. The optimum C-O-N angle (ϑ_{CON}^{0}) is set to either 115° or 155° in dependence on the distance towards the actual C-O-N angle (ϑ_{CON}). For angles above 135°, the optimum is set to 155°, while ϑ_{CON}^{0} is set to 115° for angles below 135°. The weight (ε) of the hydrogen bond is 100.

We finally chose the hydrogen bond criterion applied by Dahiyat and co-workers (Equations 3.8 and 3.9) for the assignment of hydrogen-bonded turns, but adjusted some of the parameters. Hydrogen bond energies were calculated if the acceptor-donor distance was between 2.5 and 3.5 Å (3.3 Å in the original publication). The angle defined by the backbone nitrogen, hydrogen, and oxygen atom had to be at least 130° (135° in the original publication).

Analysis of Turn Structures in Proteins

In this work, turns of length $n = x+1$ are distinguished by their class as illustrated in Figure 3.3.1: (a) normal turns are defined by a hydrogen bond interaction between the backbone carbonyl oxygen atom of residue *i* and the backbone nitrogen atom of residue *i+x*, (b) reverse turns are characterised by a hydrogen bond interaction between the backbone nitrogen atom of residue *i* and the carbonyl oxygen atom of residue *i+x*, (c) open turns show a $C\alpha_i$-$C\alpha_{i+x}$ distance between 4 and 8 Å.

Open turns that are completely spanned by or included in either reverse or normal turns were not additionally assigned. This was done to exclude any redundancy within the datasets of hydrogen-bonded (normal and reverse) and open turn structures for classification purposes.

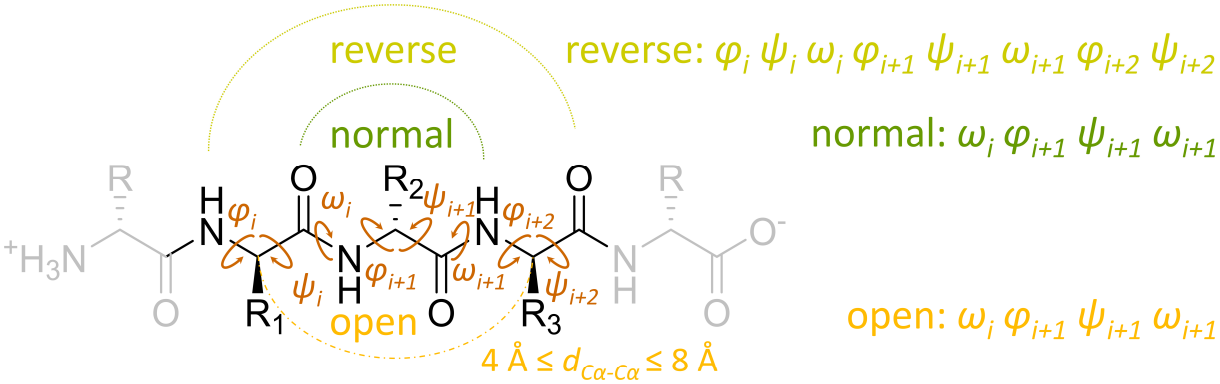


Figure 3.3.1. Turn classes analysed in this study and the underlying dihedral angles for a clustering based on ESOMs. Examples of normal turns, reverse turns, and open turn structures with $C\alpha_i$-$C\alpha_{i+x}$ distances between 4 and 8 Å are presented. They were used to derive an improved secondary structure assignment methodology. The turn categories are exemplified by a turn of three residues.

The identified turns were used for a classification based on the backbone dihedral angles of the underlying residues. In the case of alternate locations, the highest occupied one was chosen. For occupancies of 0.5, the first conformation was chosen. Some modified residues were not included and instead, a gap was introduced at their position, as they do not conform with the typical protein backbone composition (see List A3.1 for a complete list of the residues).

ESOMs are used to distinguish between turn classes with distinct conformations. The Databionic ESOM Tools[289] were applied to differentiate between turn types based on the dihedral angles of their backbone. The input dihedral angles and training parameters for the turn types are provided in Table 3.3.1.

Linear distances were used for the training of the ESOMs for the different turn categories. The backbone dihedral angles range from -180° to 180°. While the high distance between 0° and 180° is reflected by the high numerical distance of 180°, the distance between -180° and 180° is 0 although the numerical distance is 360°. This distance differences in angular space are not reflected by the numerical data originally used for the ESOM training. The applied distance function does not appropriately represent this angular space. Therefore, each dihedral angle was represented by two transformed angular values which were obtained as described in the next paragraph.

Table 3.3.1. Turn categories, their angles, and the ESOM training settings for the assignment of turn types. The number of transformed angles used for ESOM training is twice as large as the number of angles of the turn structures. The multiplier was adjusted according to the size of the respective dataset. The map for turns of the category reverse 2 was trained using a larger dataset.

Category (Class and Length)	Angles	(|Angles|)	|Turns|	Multiplier	|Neurons|	Dimensions (Rows x Columns)	Initial Radius
Reverse 2	$\psi_i, \varphi_i, \omega_i, \psi_{i+1}, \varphi_{i+1}$	(5)	977	100	97,700	382 x 256	128
Reverse 3	$\psi_i, \varphi_i, \omega_i, \ldots, \varphi_{i+2}$	(8)	134	100	13,400	142 x 95	48
Reverse 4	$\psi_i, \varphi_i, \omega_i, \ldots, \varphi_{i+3}$	(11)	4,113	100	411,300	785 x 524	262
Reverse 5	$\psi_i, \varphi_i, \omega_i, \ldots, \varphi_{i+4}$	(14)	2,892	100	289,200	658 x 440	220
Reverse 6	$\psi_i, \varphi_i, \omega_i, \ldots, \varphi_{i+5}$	(17)	1,869	100	186,900	530 x 353	177
Normal 3	$\omega_i, \psi_{i+1}, \varphi_{i+1}, \omega_{i+1}$	(4)	2,190	100	219,000	572 x 383	192
Normal 4	$\omega_i, \psi_{i+1}, \varphi_{i+1}, \ldots, \omega_{i+2}$	(7)	83,212	10	832,120	1,117 x 745	373
Normal 5	$\omega_i, \psi_{i+1}, \varphi_{i+1}, \ldots, \omega_{i+3}$	(10)	222,529	5	1,112,645	1,291 x 862	431
Normal 6	$\omega_i, \psi_{i+1}, \varphi_{i+1}, \ldots, \omega_{i+4}$	(13)	10,964	100	1,096,400	1,283 x 855	428
Open 4	$\omega_i, \psi_{i+1}, \varphi_{i+1}, \ldots, \omega_{i+2}$	(7)	99,883	1	99,883	386 x 259	130
Open 5	$\omega_i, \psi_{i+1}, \varphi_{i+1}, \ldots, \omega_{i+3}$	(10)	117,894	1	117,894	420 x 281	142
Open 6	$\omega_i, \psi_{i+1}, \varphi_{i+1}, \ldots, \omega_{i+4}$	(13)	83,711	1	83,711	354 x 237	119

The first transformed angle, ω_1, is defined as follows

$$\omega_1 = f(\omega) = \begin{cases} \omega_1 = \omega + 180^\circ \; ; -180^\circ \le \omega \le 0^\circ \\ \omega_1 = -\omega + 180^\circ ; \; 0^\circ < \omega \le 180^\circ \end{cases} \tag{3.12}$$

The second transformed angle, ω_2, is defined by the equation

$$\omega_2 = f(\omega) = \begin{cases} \omega_2 = -\omega - 90^\circ \; ; -180^\circ \le \omega \le -90^\circ \\ \omega_2 = \omega + 90^\circ \; ; \; -90^\circ < \omega \le 90^\circ \\ \omega_2 = -\omega + 270^\circ ; \; 90^\circ < \omega \le 180^\circ \end{cases} \tag{3.13}$$

The use of both transformed angles as representatives of one dihedral angle enables the calculation of realistic distances between dihedral angles as illustrated in Figure 3.3.2. In the case of geometrically very similar, but numerically distant dihedral angles, both transformed angles show a small numerical distance as compared to the underlying dihedral angles (Figure 3.3.2 (A)). For numerically and geometrically dissimilar angles, one transformed angle reflects the high dissimilarity, while the second angle indicates a low numerical distance (Figure 3.3.2 (B) and (D)). Figure 3.3.2 (C) demonstrates that numerically and geometrically similar angles are also characterised by similar transformed angles. Tse described transformation steps led to the doubling of the angles used for ESOM training (see Table 3.3.1 for the number of angles per turn category).

The visual inspection of the resulting ESOMs, the clustering according to Ultsch,[289] and the PyMOL-based visual inspection of the turns underlying the different types based on the U-Matrices verified the suitability of the applied transformation steps for the clustering of turns with respect to their dihedral angles.

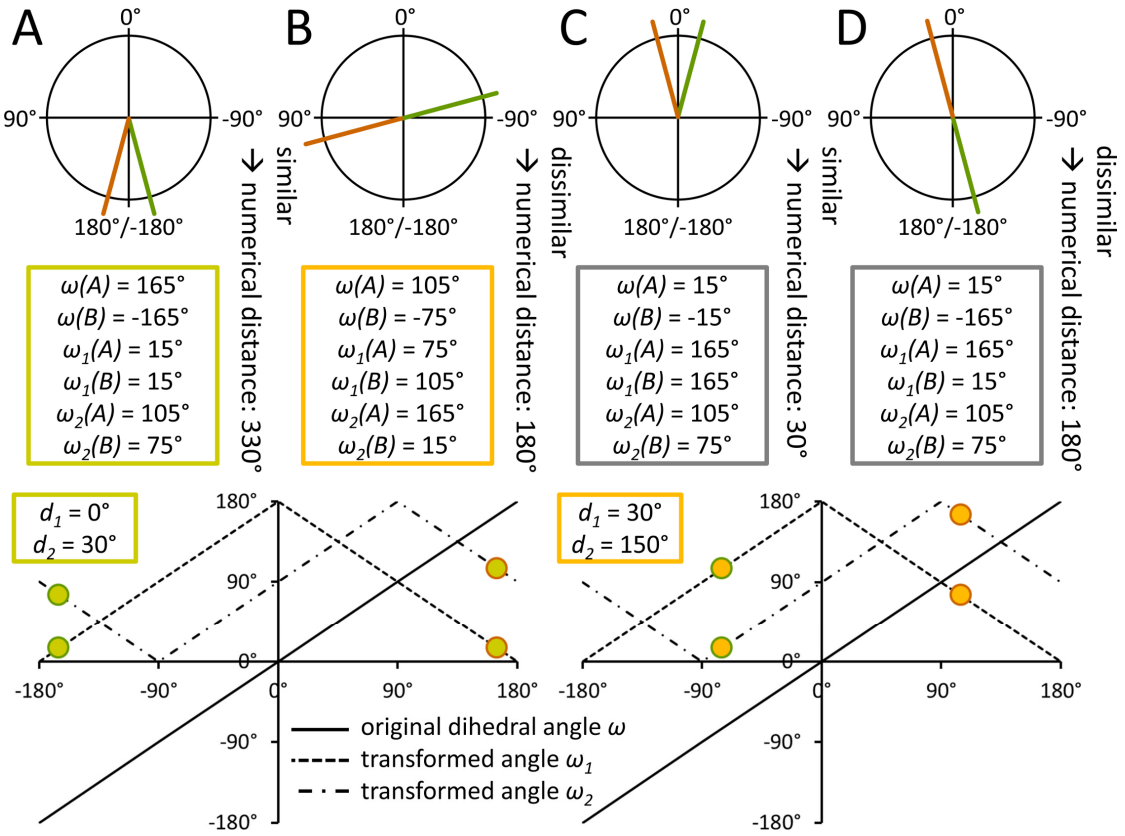


Figure 3.3.2. Transformation steps to reflect meaningful dihedral angle distances for the subsequent ESOM training. The dihedral angles of the turns A and B are depicted with the help of red (*ω(A)*) and green (*ω(B)*) lines. The diagrams at the bottom present the original angle (solid line) and the corresponding transformed angles (dotted and dash-dotted lines). The circles mark the transformed dihedral angles for cases A (left diagram) and B (right diagram). The corresponding distances between both angles for the transformed values are given.

We undertook no further pre-processing steps due to the type of the angular input data. As all analysed backbone dihedral angles consistently range between -180° and 180°, there was no need for further normalisation steps. The application of mean centring or normalisation to unit variance did not significantly influence the resulting U-Matrices after ESOM training.

Clustering of Turn Classes

The training procedure for the ESOMs was performed with the following parameters. The number of neurons was adjusted to the number of turns per category, but it was always set to at least 4,000. For especially large categories, the multiplier was decreased. The ratio of rows to columns (dimensions) was adjusted to approximately 1.5. A toroid borderless map grid was used for the clustering with Euclidean grid distances ("-g toroid", "-d euc"). The weights were initialised by sampling the hyperplane spanned by the first two principal components of the data ("-i pca"). The starting value for the radius of the neighbourhood of each neuron, which is used for the map updating, was set to approximately half the number of columns in the map ("-rs", see Table 3.3.1 for the radii which were used for the individual turn categories). A Gaussian bell-shaped kernel was used for the neighbourhood kernel function ("-n gauss"). Altogether, 200 training epochs were set for the weight adjustments ("-e 200"). The resulting weights (WTS) and U-Matrix (UMX) files were used for all further manually performed assignment steps.

For each turn type, the resulting U-Matrix was visualised using the Databionic ESOM Tools.[279] The visualisation settings differed for the open turns (Colour: 70, Clip: 60) and the turns defined by backbone hydrogen bonds (Colour: 70, Clip: 70). Different settings were initially tested to ensure a clear differentiation between turns of different conformations. Clusters of turns with similar dihedral angles and similar conformations were identified and CMX files (class masks) were generated by labelling neurons with the underlying class. This manual class assignment was accomplished considering the characteristics of a U-Matrix as defined by Ultsch[289]: (a) the best matches are positioned according to the topology of the input space (inherited from the ESOM), (b) large U-heights are assigned to neurons whose vectors are very different from the remaining ones ("mountain ranges" which hint at cluster boundaries), (c) small U-heights are indicative of neurons with weight vectors surrounded by others in data space ("valleys" which suggest cluster centres), (d) "funnels" point towards outliers in data space. The WTS files were used to project all data points to the map and the respective types were assigned by the manual generation of CMX files.

The validity of the chosen clusters was subsequently verified by visual inspection. The turns of the manually assigned types were aligned with respect to the backbone atoms and visually inspected using a PyMOL[227] script. In an iterative process, the clustering and visualisation steps were repeated until a satisfactory differentiation between types of turns with distinct backbone geometries was obtained. Outliers were excluded as far as they were recognisable in the U-Matrix. The resulting CMX and WTS files finally enable the type assignment for novel turns in protein structures.

Statistical Analysis of the Dihedral Angles, Hydrogen Bond Energies, and Distances

The mean $\bar{x}$, variance s^2, and standard deviation s of the hydrogen bond energies and Cα-Cα distances of the turns within the clustered turn types were calculated according to Equations 3.14 and 3.15.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.14}$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n} \quad and \quad s = \sqrt{s^2} \tag{3.15}$$

The calculation of statistical parameters for the dihedral angles of the assigned turn types was achieved according to the equations introduced by Batschelet.[290] For a given (untransformed) dihedral angle ω in the range from -180° to 180°, the circular average $\bar{\omega}$, variance V, and standard deviation v were calculated according to Equations 3.16 to 3.20, respectively.

$$\bar{\omega} = \tan^{-1}(S,C) = \begin{cases} \tan^{-1}(S/C) & ;\ C > S \text{ and } S \neq 0 \\ \tan^{-1}(S/C) + \pi \cdot \mathrm{sgn}(S) & ;\ C < 0 \text{ and } S \neq 0 \\ \pi/2 \cdot \mathrm{sgn}(S) & ;\ C = 0 \text{ and } S \neq 0 \\ \pi/2 \cdot [\mathrm{sgn}(C) - 1] & ;\ C \neq 0 \text{ and } S = 0 \end{cases} \tag{3.16}$$

$$C = \sum_{i=1}^{n} \cos \omega_i \quad and \quad S = \sum_{i=1}^{n} \sin \omega_i \tag{3.17}$$

$$V = 1 - \overline{R} \tag{3.18}$$

$$R = \sqrt{(C^2 + S^2)} \tag{3.19}$$

$$v = \sqrt{-2 \ln(\overline{R})} \tag{3.20}$$

Analysis of Residue Propensities (The Conformational Parameter)

The conformational parameters of the 20 standard residues and all non-standard residues were calculated within all identified turn types according to Chou and Fasman.[291] To this end, the relative frequency of residue type *i* within the complete dataset of protein structures used for the ESOM training was determined according to Equation 3.21 where n_i denotes the sum of all residues of type *i* in the dataset and $\sum_{x=1}^{20} n_x$ the number of all residues in the dataset.

$$p_i = \frac{n_i}{\sum_{x=1}^{20} n_x} \tag{3.21}$$

Accordingly, the probability that residue type *i* is found at position *a* in a certain turn of type *t* was calculated using Equation 3.22 with $n_{i,t(a)}$ as number of residue type *i* in all turns of type *t* at position *a* and $\sum_{x=1}^{20} n_{x,t(a)}$ as total number of residues in type *t* turn at position *a*.

$$p_{i,t(a)} = \frac{n_{i,t(a)}}{\sum_{x=1}^{20} n_{x,t(a)}} \tag{3.22}$$

The frequency that residue type *i* occurs at position *t(a)* ($f_{i,t(a)}$) is described by Equation 3.23, while the average frequency of finding any residue in turn type *t* at the position *a* ($\overline{f_{t(a)}}$) is calculated by Equation 3.24.

$$f_{i,t(a)} = \frac{n_{i,t(a)}}{n_i} \tag{3.23}$$

$$\overline{f_{t(a)}} = \frac{\sum_{x=1}^{20} f_{x,t(a)}}{20} = \frac{\sum_{x=1}^{20} n_{x,t(a)}}{\sum_{x=1}^{20} n_x} \tag{3.24}$$

A normalisation of $f_{i,t(a)}$ by the average frequency of finding any residue at the position *a* of turn type *t* results in the conformational parameter $P_{i,t(a)}$ (Equation 3.25).

$$P_{i,t(a)} = \frac{f_{i,t(a)}}{\overline{f_{t(a)}}} \tag{3.25}$$

Consequently, the conformational parameter of the residue type *i* in any turn type at the position *a* (*t(a)*) was calculated according to

$$P_{i,t(a)} = \frac{p_{i,t(a)}}{p_i} = \frac{n_{i,t(a)} \cdot \sum_{x=1}^{20} n_x}{\sum_{x=1}^{20} n_{x,t(a)} \cdot n_i} \tag{3.26}$$

The d-test of Wilmot and Thornton[292] was applied to assess the statistical significance of an over- or underrepresentation of residue type *i* at the position *a* of turn type *t*. The absolute value of the parameter *d* for a two-tailed test assumes values greater than or equal to 1.97, 2.57, and 3.3 for significance levels of 5%, 1%, and 0.1%, respectively. It is determined by Equation 3.27 where $n_{t(a)}$ is the number of turn type *t* in the complete dataset.

$$d = \frac{n_{i,t(a)} - p_i \cdot n_{t(a)}}{\sqrt{p_i \cdot n_{t(a)} \cdot (1 - p_i)}} \tag{3.27}$$

The residue type *i* was assigned as being statistically overrepresented if the value of *d* was larger than or equal to 1.97, while values smaller than or equal to -1.97 indicate an underrepresented residue type. The significance levels were incorporated by colour coding in Tables A3.1-4, A3.6, A3.9, A3.13-16, A3.18, and A3.19.

3.3.2 Analysis of the Turn-Based Secondary Structure Classification Method SCOT

Details regarding the development of the SCOT software for the automated assignment of SSEs in protein chains will be an integral part of the doctoral thesis of Dipl.-Inf. Tobias Brinkjost who implemented the software SCOT. The general approaches underlying the methodology are introduced and discussed in Section 4.3.3.

The SSE assignments with SCOT were analysed with respect to geometric characteristics according to the parameters twist, rise, radius, and bending angle as introduced in the publication of Kumar and Bansal.[122] The scaled B-factor for all protein backbone atoms as described by Carugo and Argos[293] was calculated to evaluate the intrinsic flexibility of the assigned SSEs. This scaling to a mean of zero and unit variance leads to the comparability of the B-factors of all protein structures in the dataset.

The consensus of different SSAMs for the assignment of β-strands, as well as left-handed and PPII helices, was calculated by extracting a sequence-based binary fingerprint (1 for residue positions for which the identical SSE type was assigned, 0 for all other positions). The Tanimoto coefficient for one structure with SSEs assigned by two SSAMs was calculated. The reported consensus values correspond to the mean Tanimoto coefficients for the complete dataset.

All analyses were performed for the dataset that was also used for the training of the ESOMs. It comprises 3,597 diverse protein chains (see Section 3.3.1).

3.4 Trypanothione Synthetase

3.4.1 Comparative Modelling

The initial model of *Lm*TryS (in the presence and absence of different substrates) was obtained by the modelling of missing sequence regions of the structure with the PDB-ID 2vps based on the *Ec*GspS structure with the PDB-ID 2ioa.[162] The resulting model served as the template for the homology modelling of the TryS structures from *C. fasciculata*, *L. infantum*, *T. brucei brucei*, and *T. cruzi*. The "Homology Model" module of MOE2013[257] was used to build the structures of the enzymes according to the sequences given in Table A4.1. It was also applied to model the complete structure of the template as described previously.[162] The initial progressive sequence alignment was performed using the BLOSUM85 matrix and 1,000 iterations. The default parameters were used for the homology modelling with the following adjustments: 100 models were built with 100 side-chain samples per model. The intermediate models and the final models were refined with an RMS gradient of 0.5. In the case of the substrate-bound models, the substrates of the *Lm*TryS model were included as the environment for induced fit. Hydrogen atoms were added using the "Protonate3D" tool in MOE2013. The model quality evaluation results for all models with PROCHECK[255] are presented in Table 3.4.1. G-factors below -0.5 point towards unusual structure properties, while G-factors below -1.0 indicate highly unusual structural features.

Table 3.4.1. PROCHECK analysis results for the original X-ray structure with the PDB-ID 2vps and the homology models used in this study.

Structure	Number of Residues	G-Factors			Ramachandran Outliers		Bad Contacts
		Overall	*Dihedral Angles*	*Main-Chain Covalent Forces*	*Dis-allowed Regions*	*Generously Allowed Regions*	
2vps	603	-0.07	-0.37	0.38	0.2%	0.8%	14
*Lm*TryS	648	-0.23	-0.46	0.09	0.4%	0.4%	1
*Cf*TryS	648	-0.31	-0.60	0.09	0.5%	0.7%	2
*Li*TryS	648	-0.27	-0.54	0.11	0.5%	0.5%	1
*Tb*TryS	626	-0.36	-0.64	0.03	0.6%	1.5%	1
*Tc*TryS	634	-0.30	-0.65	-0.01	0.5%	1.3%	1

3.4.2 MD Simulations

System Preparation

The MD simulations of all models were conducted with AMBER14.[236] Sander was used for the minimisation and equilibration steps and the GPU-accelerated version of PMEMD (Particle Mesh Ewald Molecular Dynamics)[294,295] for the subsequent simulations. Parameters for ATP were adapted from previous studies.[296] The parameterisation of GSP and GSH was realised using Antechamber and GAFF (the General AMBER Force Field). Partial charges were assigned using the semi-empirical method sqm.[297] The same holds true for the compounds MOL2008 and FS-554. The ff14SB force field[298] was applied to describe the protein residue atoms. The structures were placed in an octahedral TIP3P water box whose edges extend by least 12 Å in each direction from the solute molecules. Counter-ions (Na^+ or Cl^-) were added to obtain a charge neutral system with monovalent ion parameters for TIP3P water.[299]

Simulation Steps

The minimisation and equilibration steps were performed with SANDER. A two-step energy minimisation was followed for all prepared systems. First, 50 minimisation steps with steepest descent followed by 200 steps of conjugate gradient were performed using harmonic positional restraints of 25 kcal/(mol·Å²) for all solute atoms. Second, these steps were repeated with reduced restraints of 5 kcal/(mol·Å²). Afterwards, 50 ps of NVT simulation runs were performed to incrementally increase the system temperature from 100 K to 300 K using the Berendsen thermostat. The subsequent NPT simulation for 150 ps at a constant pressure of 1 atm (using the Berendsen barostat) was run to adjust the volume of the unit cell. Finally, the positional restraints were incrementally reduced by 1 kcal/(mol·Å²) in 50 ps steps of NVT simulations. The last steps of the NVT simulation (50 ps) were performed without restraints.

The following 600 ns NVT simulation runs with PMEMD on GPU were performed with a step size of 2 fs and snapshots were saved every 20 ps. A constant temperature of 300 K was ensured using the Berendsen weak coupling algorithm. The SHAKE algorithm was applied to constrain the bond lengths between hydrogen atoms and their covalent partners. The PME (Particle Mesh Ewald) conditions were applied for the treatment of long-range electrostatics. An 8 Å cut-off distance was set for the calculation of short-range non-bonded interactions.

Analysis

CPPTRAJ[300] of the AMBER14 Tools package was used for the analysis of the resulting MD trajectories (clustering, analysis of RMSD and RMSF (root-mean-square fluctuation) values, distance measurements). The hierarchical agglomerative approach with average linkage was applied for an RMSD-based clustering of the snapshots. The clustering was performed focusing on the residues in a 5 Å radius of the TryS ATP binding site. The minimum distance between the clusters was set to be larger than 1 Å.

Visualisations and further analyses were realised in VMD[231] and UCSF Chimera.[228] The hydrogen bond analysis in VMD was performed with a distance threshold of 3.5 Å and a maximum deviation of 30° from the optimum donor-hydrogen-acceptor angle.

3.4.3 Paullone Binding Mode Prediction

The compounds MOL2008 and FS-554 (see Figure A4.1 for their molecular structures) were docked into the ATP binding site of TryS representatives of the clustered MD snapshots (see Section 3.4.2). Although GOLD (Genetic Optimization for Ligand Docking) offers the possibility to use a ring conformation library, no alternative ring conformations were generated when docking the compounds with a paullone scaffold. Therefore, the molecules were separately docked in two different ring conformations depicted in Figure 3.4.1. These conformations were generated based on the two crystal structures of the seven-membered ring as extracted from the CSD (IDs: POHWIT, POHWAL).

The docking with two ring conformations of both molecules was performed with GOLD (ensemble docking with multiple TryS representatives). A search efficiency of 100% was used and 100 GA runs per molecule were performed. The early termination option was disabled.

The molecules ATP, GSH, and GSP of the snapshots from the MD with all substrates of the second reaction step were used to define the binding sites. Apart from that, the default settings were applied.

The resulting binding modes were visually inspected, and a distinct set of probable binding poses was chosen for MD simulation analyses (see above for details on the MD simulations).

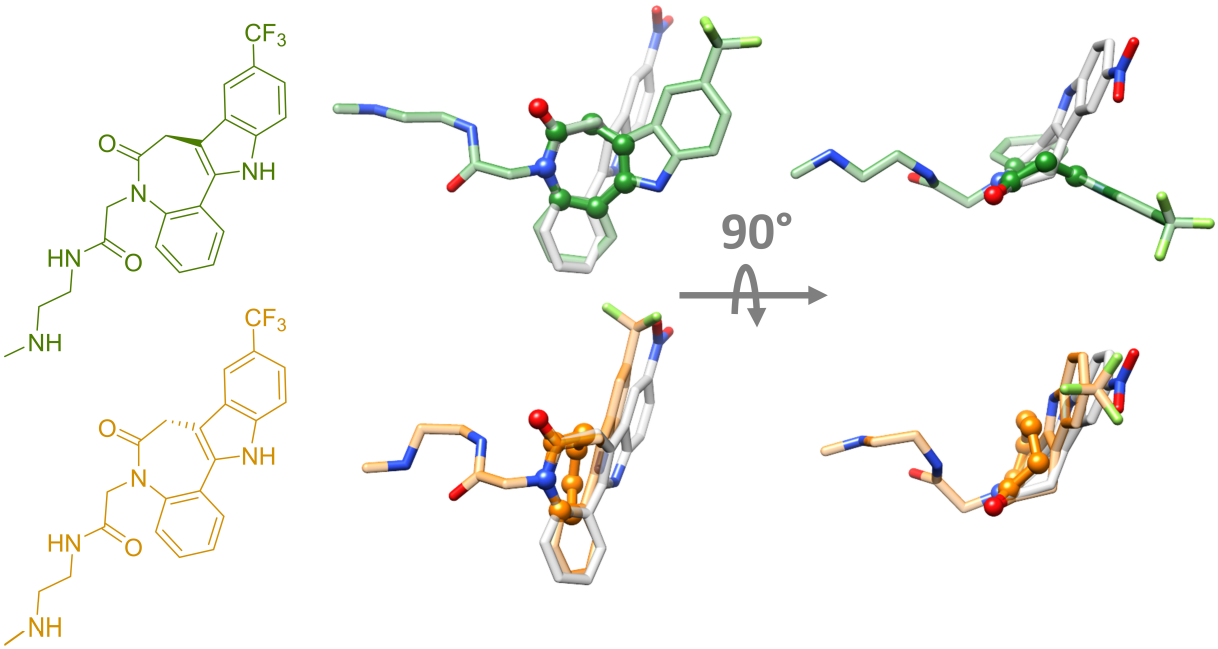


Figure 3.4.1. The two known conformations of the seven-membered ring of the paullone scaffold. The conformation of FS-554 corresponding to the crystallized structure with the CSD-ID POWAHL is given in orange (CSD structure conformation in ball-and-stick representation). This conformation also corresponds to that in the seven-membered ring of the alsterpaullone in GSK-3β (PDB-ID 1q3w, grey). The second possible ring conformation for FS-554 is given in green with the corresponding structure from the CSD (POHWIT) in ball-and-stick representation. For comparison purposes, the conformation of the alsterpaullone in GSK-3β is also given.

3.4.4 Comparison to Protein Kinases Based on the Secondary Structure Elements

The methods SLOT and LOCK2[301] were applied to compare the secondary structure arrangements in the proximity of the ATP binding site in the structures of *Lm*TryS and *Li*TryS and the binding sites of a large set of protein kinases in complex with a main pocket-binding ligand (see below for more information about the dataset). To this end, SSEs were assigned for all dataset structures using the SCOT methodology (see also Section 4.4). Helices and strands were split using the positions of the assigned kink residues.

Graph-Based Comparison of Secondary Structure Element Arrangements (SLOT)

Our in-house method SLOT was applied for the graph-based comparison of SSE arrangements. It was developed in close collaboration with Dipl.-Inf. Tobias Brinkjost who also implemented the software. In short, the arrangement of helices and strands is modelled using complete undirected graphs, i.e., by vertices corresponding to the SSEs, while their spatial relationship is encoded by the edges (without direction) that connect all vertices in a pairwise manner.

For each SSE, a vertex is added to the graph and is labelled with segmentation points. These points enable a modelling of the overall geometry of an SSE (especially for bent instead of linear helices and strands). They are differently placed for helices and strands (see also Figure 3.4.2). In the case of helices, the centre of each helix turn is assigned a segmentation point. Segmentation points in strands are placed at the centre of the line connecting the backbone nitrogen atoms of two consecutive strand residues.

The vertices are additionally labelled with the SSE class (distinguishing between β-strand, α-helix, 3_{10}-helix, and π-helix). Each vertex is connected to all other vertices by edges (complete graph) which are labelled with a distance matrix. Each distance matrix contains all pairwise distances of the segmentation points of one vertex to the segmentation points of the other vertex. A minimum of two segmentation points is necessary for the modelling of SSEs, i.e., strands with less than two residues and helices with less than two helical turns are not modelled. It is worthwhile mentioning that the segmentation points at the vertices are only used for the creation of the distance matrices and are not considered in the comparison or matching procedure.

The final graphs for several TryS structures (queries) and protein kinase structures (targets) were compared to find the maximum common subgraph (MCS) for each query and target structure pair. The MCS of two graphs is defined as the induced subgraph which maximises the number of matching or compatible vertices in both graphs. Each MCS represents a maximal possible match consisting of the matched vertices and edges which represent the similarly arranged SSEs sharing a common geometry and class in both proteins.

Different parameters were optimised which will be summarised in the following. The vertices of two graphs are not matched if they represent different SSE classes or the difference in the number of compared segmentation points exceeds 50% of the number of the vertex with the larger number of segmentation points. For two edges, each connecting a pair of vertices of the two graphs, the maximum accepted difference of the distances between the terminal segmentation points (outer distances) that are matched was set to the static distance difference of 1 Å plus the dynamic distance difference (10% of the shorter distance). The maximum accepted cross distance difference between the terminal segmentation points was set to 2 Å plus the dynamic distance difference (10% of the shorter distance). The matrix comparison starts at the top left of the query matrix. A set of outer and cross distances whose differences are below the predefined thresholds is searched for by comparing query matrix subsets to the target matrix. If a match between the minimum number of required segmentation points (2 in this analysis) is found, the matrix comparison stops. Although other possibilities for the segmentation point matching might lead to smaller distance differences, the distance matrix comparison ends whenever such a match is found.

In addition to the distance thresholds, an angle threshold was set. For the matched segmentation points of the two SSEs in the query and target proteins, the angle between the vectors connecting the terminal matched segmentation points is calculated. The angle difference between the angles of the vector pairs of both proteins should not exceed 35°.

If both criteria (distance and angle thresholds) are met, both vertices and the connecting edge are compatible. Importantly, the disregard of other potentially better segmentation point matches has no impact on the final MCS. Figure 3.4.2 gives an example of two simplified input protein structures consisting of only two SSEs. The matched outer distances and cross distances are highlighted. The parameters were optimised based on the comparison of the representative TryS structures to GSK-3β. The parameters which led to the anticipated alignment of both ATP binding sites were chosen.

Each of the MCSs (of equal size independent of the types of matched SSEs) for the two models under investigation is a possible structural match, given the aforementioned thresholds. The number of cliques, i.e. the subset of vertices in which every two distinct vertices are adjacent, can therefore vary, depending on the protein pair of interest. The number of cliques corresponds to the number of MCSs per structural pair. A score is calculated based on the number of similarly arranged SSEs and is identical for all MCSs. In other words, multiple similar arrangements of SSEs can be found for a pair of protein structures.

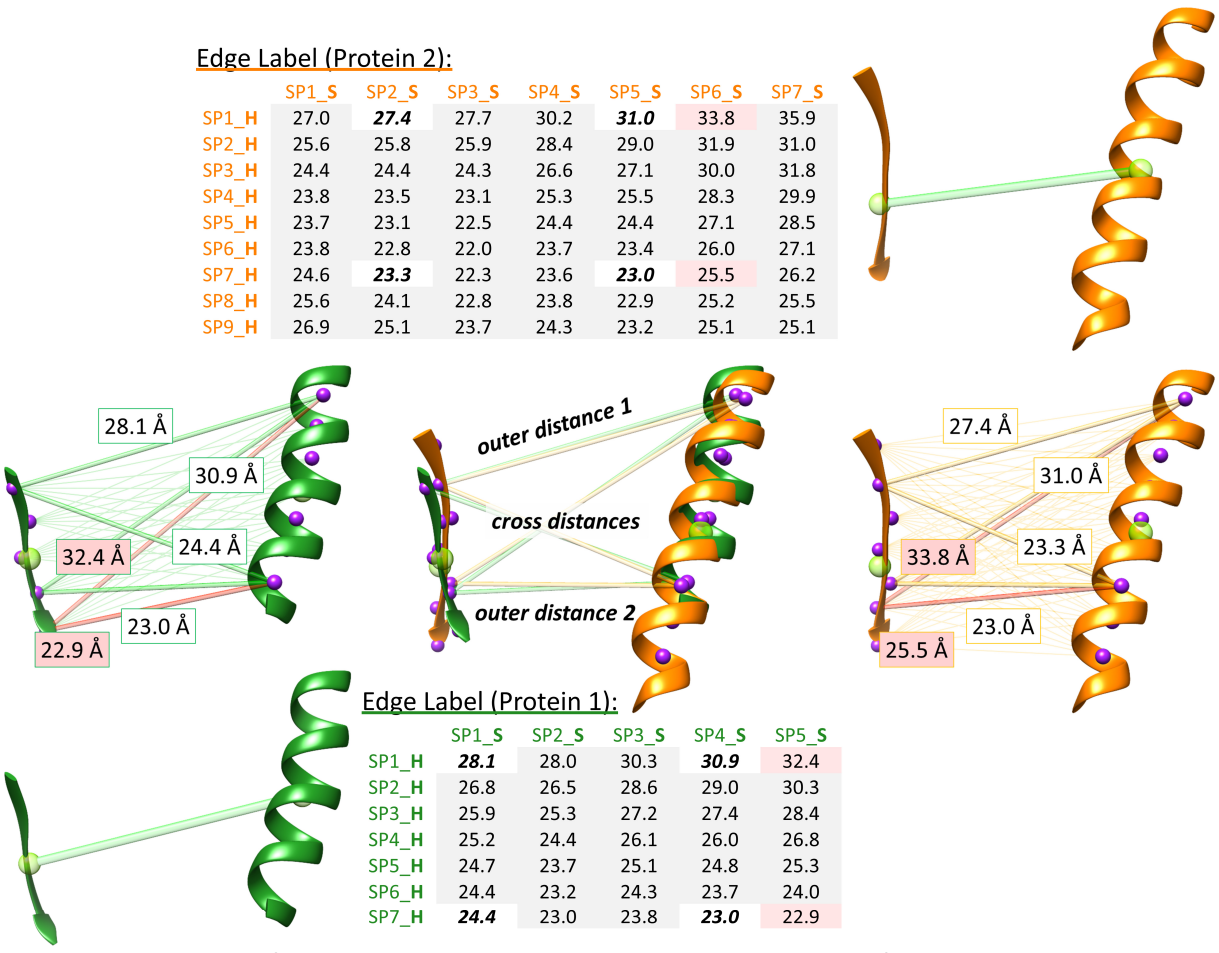


Figure 3.4.2. Example for the distance matching procedure used to identify similar SSE arrangements with SLOT. Purple spheres represent the segmentation points used for the distance matrix calculation. Green spheres represent the vertices for the respective SSEs which are connected by edges (green lines) labelled with the distance matrices. The matched outer and cross distances (thick lines) of both proteins are highlighted in white within the complete distance matrices of both edges (thin lines). One non-matching outer distance with the corresponding cross distance is additionally highlighted in red. In the centre of the figure, the final overlay for this matching procedure is given in the centre.

These modelling and comparison procedures were performed for the helices and strands which are part of the ATP binding sites (the complete SSEs in a 5 Å neighbourhood of the ligand atoms) and the complete chains. In the latter case, only the synthetase domain of TryS was used for comparison purposes.

Vector-Based Comparison of Secondary Structure Element Arrangements (LOCK2)

The SSE arrangements in the chains and the binding sites of TryS and protein kinases were additionally compared with LOCK2.[301] This vector-based alignment method was applied with the default settings. It only distinguishes between helical and strand conformations.

3.4.5 Binding Site Comparison

Dataset Preparation

Protein kinase structures in complex with main pocket binders were downloaded from the new KLIFS (Kinase-Ligand Interaction Fingerprints and Structures) database (11/2017).[302] The dataset was prepared as described for the binding site comparison benchmark datasets in Section 3.2.1. Some ligands with identical ligand IDs were present in two differently localised binding sites (e.g., for p38α). For these structures, a manual selection of the main pocket ligand was performed.

The *Lm*TryS and *Li*TryS structure snapshots of the MD simulations in the presence of ATP or all substrates of the second $T(SH)_2$ biosynthesis step were clustered with respect to the RMSD of all residues in a 5 Å radius of the bound ATP molecule (see Section 3.4.2 for more details). Representatives of clusters with occupancies above 10% were chosen for the binding site comparisons. This filtering step led to nine *Lm*TryS and four *Li*TryS representative structures.

Binding Site Comparison

The following binding site comparison tools were used to compare all *Lm*TryS and *Li*TryS representatives against structures extracted from the KLIFS database: Cavbase, IsoMIF, ProBiS, Shaper (with MOL2 files as input), SiteAlign, SiteEngine, SiteHopper, SMAP, and TM-align. FuzCav, PocketMatch, and RAPMAD were excluded in the following steps as they do not provide binding site alignments. The results of TIFP and Grim were neglected due to the potential bias to the interactions between TryS and the complexed ATP molecule. The parameters described in Section 3.2.4 were applied for all methods.

The results were filtered according to predefined thresholds based on dataset 1 of Section 3.2.1. The thresholds were chosen according to the upper whisker (for similarity scores) or lower whisker (in the case of distances measures) for the inactive pairs of this dataset. Consequently, the thresholds were set to > 8.35, > 55.39, < 0.14, > 0.33, > 0.82, > 764.3, > 0, > 0.96, and > 0.47 for Cavbase, SMAP, SiteAlign, TM-align, SiteHopper, SiteEngine, ProBiS, Shaper, and IsoMIF, respectively. Matches of the protein kinase binding sites to any of the representative structures of *Lm*TryS and *Li*TryS, respectively, were used as detected binding site similarity pair. However, the results for both proteins were analysed separately and a consensus approach was pursued.

For each TryS-kinase pair, the number of tools that scored this match above (for similarity scores) or below (for distance scores) the given threshold was calculated. Protein kinase-TryS matches which were found by at least five (for the five *Li*TryS structures) or six (for the nine *Lm*TryS structures) methods were analysed. The SiteHopper scoring measures PatchScore, ColorTanimoto, and ShapeTanimoto were extracted for the filtered consensus pairs with similar binding sites to identify the most similar ATP pockets. To this end, the matches between the protein kinase and the TryS representative with the highest PatchScore were extracted. Afterwards, the corresponding SiteHopper matches were visualised with VIDA to finally decide on the most appropriate protein kinases for a repurposing strategy.

The similarities of PKACA (cAMP-dependent protein kinase catalytic subunit α) and Erk2 to TryS were the most convincing ones. Ligands of known complex structures of PKACA and Erk2 were downloaded from the PDB and used for molecular docking into the ATP binding site of the TryS representative structures.

3.4.6 Docking of the Protein Kinase Similarity-Based Compound Library

Selection of Compounds for Molecular Docking

For the most similar kinases (PKACA, Erk2), the bioactivity data of the ChEMBL[303] database (11/2017) were downloaded and prepared. The preparation steps included the removal of bioactivities with a confidence score below 9. Compounds with IC_{50} or K_i values larger than or equal to 10 µM were excluded. Small molecules with bioactivity values with ">" relations were removed. As the inhibitor concentrations vary for the activity types "% inhibition" and "residual activity", these bioactivities were not considered.

The Tanimoto similarities of the filtered compounds to all commercially available compounds of the MolPort database (11/2017) were calculated based on the ECFP-like Morgan[304] fingerprints which were calculated with RDKit[305]. Compounds showing a Tanimoto similarity of at least 0.9 to any compound inhibiting PKACA or Erk2 were used for the subsequent docking studies (32 compounds). Moreover, the Murcko scaffolds of the known inhibitors were extracted in KNIME[261] and a substructure search was performed for all scaffolds with a molecular weight of at least 150 g/mol. The resulting hits were reduced using cut-offs for the molecular weight (< 600 g/mol) and the number of aromatic rings (< 5). Additionally, the compounds with unconstrained chirality were excluded. The remaining 240,365 compounds and the 32 compounds from the similarity search were prepared by adding hydrogen atoms with MOE ("Wash"). A subsequent energy minimisation (using the MMFF94x force field) was performed to generate the initial 3D conformations for the molecules of the library.

Molecular Docking

All representative protein structures which were also used for the binding site comparison analyses were prepared as follows. The structures were aligned with respect to the ATP binding site and protonated with the "Protonate3D" tool as implemented in MOE[257] using the Amber10:EHT force field. The prepared structures were saved as PDB files.

The bound ATP molecule of the first representative structure was used for the binding site definition with a 6 Å radius (the flood fill cavity detection option was disabled). After the initial docking of small molecules whose binding mode is known from PKACA and Erk2 complex structures, the docking parameters were additionally adjusted. A protein hydrogen bond constraint was introduced to penalise binding modes that do not show a hydrogen bond towards residue Phe586. Finally, 100 GA runs per molecule with a search efficiency of 1 were performed. The force field-based scoring function GoldScore was used for pose prediction.

For the docking of the larger dataset with 240,365 compounds, the initial docking parameters were additionally adjusted in favour of runtime. Only ensemble structures that were preferentially used for docking of the known kinase inhibitors were included for the final docking. Moreover, the search efficiency was reduced to 0.25. Only 25 GA runs per molecule were performed. These adjustments did not lead to significantly worse results when applied for the docking of known inhibitors.

4 RESULTS AND DISCUSSION

4.1 Binding Site Prediction

4.1.1 Motivation

The nature of binding sites is manifold. Their prediction and analysis are the very first steps of most structure-based *in silico* approaches. We find proteins binding to low molecular weight compounds, peptides, proteins, or nucleic acids. All these types of binding sites play crucial roles in drug discovery, but they are different with respect to their shape and physicochemical properties. Specialised methods were developed to predict these different binding site types. However, the analysis of druggable binding sites occupied by small molecules was the focus of this work. Therefore, this benchmark analysis is restricted to the prediction of small molecule binding sites and does not include specialised methods.

The large number of publicly available software tools to identify small molecule binding sites renders it difficult to decide on a certain method (see Table A1.1 in Appendix 1 for the complete list of published methods). The individual methods were initially tested on different and partially misleading datasets. Therefore, their comparison by means of objective and diverse datasets is indispensable. To this end, we must draw our attention towards the different aspects of binding site prediction: first, the underlying methodology used to identify cavities, and second, the scoring scheme used to rank the putative binding sites. With respect to the latter, it is frequently noted that the binding site with the largest volume of a single-chain enzyme is found to be bound to a ligand in 83% of the cases.[38] Based on this finding, most early and geometry-based prediction methods rank the detected pockets with respect to their size. Nevertheless, alternatives emerged as, for example, the ranking based on the evolutionary residue conservation.[306] Additionally, the elucidation of druggable and non-druggable binding sites[9,307] led to a general reconsideration of the binding site prediction scoring schemes. The focus was drawn to the identification of binding sites which bind to drug-like ligands leading to the desired modification of protein activity.[9] Druggability measures were included in the process of binding site prediction, e.g., for the binding site predictions generated by fpocket[8] and DoGSite (DoGSiteScorer[254]). Other tools offer again other means of ranking the identified pockets. The choice for the ranking measure can have a huge impact on the outcome of the study.

As the herein analysed standalone tools differ largely not only in the binding site ranking but also with respect to the detection algorithms, the next paragraph deals with their basic methodologies and ranking schemes.

4.1.2 Binding Site Prediction Methodologies

In the following, the underlying concepts of all applied methods are explained in detail. Exact knowledge of the basic principles and assumptions of these approaches can help to understand potential drawbacks and pitfalls. Although a number of review articles briefly summarise numerous published methodologies,[45–47] they do not cover all recent software and developments.

Three basic concepts (Figure 4.1.1) can be distinguished according to the underlying assumption which is utilised to answer the question "What makes a binding site a binding site?"[308] Geometry-based methods rely on the finding that most small molecule binding sites are characterised by a concave shape, in contrast to flat invaginations of the protein surface which are most pronounced for PPI (protein-protein interaction) interfaces. One of the first approaches that analysed this feature as binding site determinant was POCKET.[309]

Other methods were designed to search for "sticky" region on the protein surface, i.e., protein patches that show high theoretical interaction energies towards selected probe molecules. The ground-breaking application for those approaches was GRID.[284] The increase in the available computational resources led to the transition from probes towards representative sets of small molecular fragments or even small molecules. Molecular docking of these molecules can lead to the identification of energetically favourable surface regions for the binding of small molecules (e.g., with MolSite[215]).

Besides these two approaches, the enormous growth of data on protein sequences and structures, binding sites that fulfil distinct biological functions, and evolutionary relationships led to the development of knowledge-based methods. First, conservation-based methods, which solely rely on protein sequence relationships, were used to identify conserved residues as an indicator for putative binding sites (e.g., with the method Sequence Space[306] as one of the first approaches). Again, the availability of more computing power enabled not only fast MSAs, but also the fast comparison of proteins and binding sites. Consequently, our knowledge of the structures of a diverse set of protein-ligand complexes enables the prediction of novel sites using structure comparison approaches (e.g., ProBiS[61]).

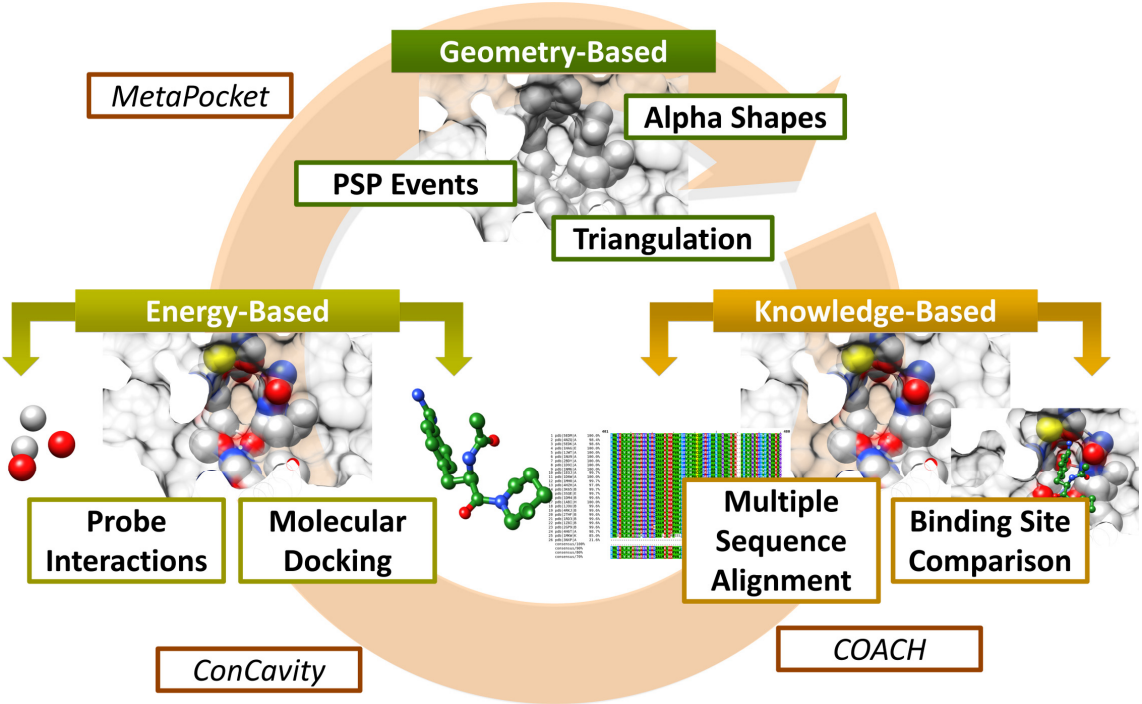


Figure 4.1.1. Overview of binding site prediction methodologies. The colours which are used in this overview will also be applied for the differentiation of geometry-based (dark green), energy-based (light green), knowledge-based (yellow), and combined approaches (red) throughout the manuscript.

Finally, the combination of the different strategies enables the emergence of the more reliable consensus or meta methods for the detection of small molecule cavities. In MetaPocket 2.0[171] and COACH,[36] the results of different prediction methods are combined. ConCavity[34] applies different geometry- and energy-based methods and additionally uses evolutionary residue conservation information for reliable pocket predictions.

In the following, the methodologies, which are analysed in more detail herein, are introduced in chronological order as several of them relate back to preceding approaches. The description is focused on the protein representation, preparation, and the scoring measure for the ranking of the putative binding sites.

Geometry-Based Binding Site Prediction Methods

VOIDOO[20] The software detects invaginations and voids in protein structures by using a methodology called “atomic fattening”. The approach is not suited to identify pockets which are solvent-exposed but cannot be closed off by atomic radii. The algorithm starts by generating a 3D grid around the molecule with a spacing of 1 Å. A flood fill algorithm is applied to exclude grid points which are connected to the outside world. Grid points which do not change their value as they are disconnected from the outside are assigned as cavities and their volume is calculated. Subsequently, the vdW radii of the protein atoms are increased (by the so-called “fattening factor”) and the flood fill algorithm is applied. New cavities are identified and after a predefined number of iterations or if the number of cavities declines, the “fattening factor” yielding the largest number of cavities is chosen. A final refinement and delineation of cavities lead to a set of putative cavities that are represented as grids and pocket centres. The predicted cavities are not ranked but annotated by their volume.

SURFNET[209] The SURFNET algorithm starts by mapping the POI onto a grid and assigning the appropriate vdW radii to all atoms of the input molecule. Afterwards, centres of spheres are iteratively set between all atom pairs of the molecule or a predefined set of atoms. The radius of these spheres is decreased until the sphere does not overlap with any vdW sphere of the protein atoms or until the radius falls below 1 Å (in this case, the sphere is not kept). Several gap spheres that fill invaginations and clefts of the protein are assigned. They are used to annotate the grid points with grid density values according to a predefined Gaussian function. A 3D surface is generated and distinct gap regions are detected which can be ranked according to a crude volume estimate as calculated based on the grid.

PASS[11] The algorithm PASS (Putative Active Sites with Spheres) relies on the so-called probe spheres. First, the protein surface is coated by a layer of probe spheres that are filtered according to different criteria. The probes that clash with protein atoms or are insufficiently buried are excluded. Subsequently, a distinct set of probe spheres is retained that ensures that two probe centres are separated by at least 1 Å. The iterative addition of sphere layers to the probes leads to the filling of the protein invaginations with probes. The active site points are assigned by searching for central probes in regions with spheres which are characterised by a high buriedness. The final scoring measure, the probe weight, is assigned with re spect to the number of adjacent probes and their degree of buriedness.

LIGSITE^{CS}[12] The LIGSITE[273] algorithm was developed as an enhancement of the original method POCKET.[309] The protein atoms are placed on a uniform grid. The grid points adjacent to any protein atom are analysed by scanning the environment in the x-, y-, and z-directions. The protein-solvent-protein events along the directions are counted for each solvent accessible grid point. In addition, these scans are extended by scans along the cubic diagonals leading to a maximum of seven protein-solvent-protein events per grid point. A binding region is defined by all adjacent grid points whose protein-solvent-protein count is above a user-defined threshold. Additionally, a threshold for the minimum number of grid points is introduced to ensure the exclusion of small protein voids. The advanced method LIGSITE^{CS}[12] (CS, Connolly Surface) considers the Connolly surface of the proteins. Instead of counting the protein-solvent-protein events, the surface-solvent-surface elements are summed up. The detected cavities are ranked by counting the number of grid points per cavity. In a further extension, the LIGSITE^{CSC}[12] (CSC, Connolly Surface, Conservation) web server, the detected pockets are scored according to the degree of conservation of the involved residues.

EPOS^{BP}[194] A reimplementation of the PASS algorithm is used in the EPOS^{BP} (Ensemble of Pockets on Protein Surfaces with BALLPass) approach for predicting transient pockets on protein surfaces using MD trajectories. The output includes the negative image of the pocket and the pocket-lining atoms (i.e., the residues within a 5 Å radius of the PASS probes). The calculation of the pocket properties volume, polarity, and depth enables a closer examination of the identified pockets. The volume is approximated by using the PASS probes and is used to rank the putative binding sites.

PocketDepth[206] A 3D grid enveloping the protein is constructed by PocketDepth. The grid cells are assigned as internal (surrounded by other internal cells on six faces), external, and surface cells (the atom-occupied cells within 3 Å). A depth-first search for connected surface cells is undertaken, i.e. the grid cells are traversed recursively along six faces from each surface cell and marked. The traversal closes if no other surface cell is encountered. The number of connected surface cells is used to identify the internal voids which are characterised by a number below 500 by default. Pairs of boundary atoms (lying within 2 to 15 Å) are connected by grid bars which do not pass through the protein. For each cell, the number of passing grid bars is calculated and defined as the depth factor. For cells that were not passed by grid bars, the depth factor is set to the average of the surrounding cells. A density-based clustering is used to identify a distinct set of pockets. Their rank corresponds to the number of site grid points, but can be adjusted using other descriptors (e.g., the depth factor).

PocketPicker[207] For the predictions, a 3D grid which covers all protein atoms is generated and the buriedness values of the grid probes in close proximity to the surface are calculated. The calculation of the buriedness is realised via an elaborate scanning along 30 directions around the grid probe resulting in a buriedness index ranging from 0 to 30. Neighbouring grid probes with high buriedness indices (16 - 26) are clustered to identify the putative binding sites. A subsequent calculation of shape descriptors offers the possibility to compare the pocket shapes. The pocket size is used to rank the putative cavities.

PHECOM[204] The method PHECOM (Probe-based HECOMI finder) exploits the definition of a pocket as an invagination of the protein where small probes, in contrast to larger ones, can enter. The probes are placed on the vdW surface of the protein. The small probes which overlap with large probe spheres are removed. Consequently, the remaining small probe spheres indicate the predicted cavities. The radius of the small probe was defined as that of a single methyl group, while the default radius of the large probe sphere was determined empirically. It became obvious that different radii for the large probe are preferable, depending on the binding site type. The incorporation of varying large probe sphere radii helped to identify different pocket types. Cavity size and depth are characterised by the radii of the small and large probe spheres. The inclusion of the $R_{inaccess}$ measure (the minimum inaccessible radius) allows for the estimation of the pocket shallowness. The predicted cavities are ranked according to the pocket volume.

VisGrid[211] The protein structure is placed onto a 3D grid. Filled, empty, and surface voxels are identified. Local geometric features of the binding sites are derived using a visibility criterion. From a target position at the protein's surface (surface voxel) the fraction of visible directions is defined (visibility). The number of possible directions using a grid representation is 26, while it rises to 98 if two layers surrounding the surface voxel are used. One direction is defined as "visible" if no filled or surface voxel is hit in up to 20 steps. Pockets are characterised by low visibility, whereas flat cavities comprise regions without protrusions or hollow regions. A search for flat circular regions without invaginations and protrusions is used to identify shallow sites. Cavity voxels with visibility smaller than or equal to 8/98 are finally merged if they are within 2 Å of each other. The number of voxels per site is used for ranking purposes.

EDTSurf[193] The method starts by a surface construction with the help of the theory of Euclidean Distance Transform (EDT). The part of the triangulated surface that is not in contact with the outside space (inner molecular surface) is annotated as void. Consequently, the algorithm discovers the buried cavities inside the macromolecules. No pocket ranking is performed, but the pocket volume is reported.

fpocket[196] Following Voronoi tessellation, α spheres at the Voronoi vertices are pruned with minimum and maximum radii. The remaining spheres are labelled with respect to the contacting atoms (apolar, polar). First, the spheres are clustered by a distance criterion. Then, single sphere clusters are removed and the centre of mass of the remaining clusters is calculated. Clusters with neighbouring centres of mass are merged. Finally, a multiple linkage clustering based on a comparison of all vertices of one cluster to all vertices of another cluster is applied to further merge adjacent clusters. The resulting pockets are ranked according to their likelihood to bind small molecules (pocket score). This score was derived from a partial least squares model which is based on different pocket descriptors (the number of α spheres, the pocket density, the polarity score, the mean local hydrophobic density, etc.). Additionally, a drug score was developed to prioritise druggable cavities (fpocket version 2.0)[8] which could also be used for the pocket ranking.

Voronoia[213] The software was designed to detect voids in protein structures, analyse protein packing, find solvent pockets within the protein, and identify the functionally important cavities in membrane channels and transporters. A Voronoi Cell algorithm is applied to identify solvent excluded voids and determine their volume. The predicted cavities are ranked according to the number of cavity neighbours.

ghecom[198] The original probe-based method PHECOM was further extended and improved to develop a novel method named ghecom. As before, a grid is generated to represent protein atoms and probes. The former probe sphere placement according to the PASS algorithm was adjusted. Grid points are annotated depending on their accessibility to small and large probe spheres. Compared to PHECOM, the algorithm shows a higher efficiency. It is independent of heuristics for the protein-contacting probe placement. The pocket volume is used for the ranking of the resulting cavities.

DoGSite[191] A grid representation of the protein with grid points labelled with free or occupied is the basis for pocket detection. Invaginations of the protein surface are detected with the help of a Gaussian filter with different radii by means of the 3D Difference of Gaussians (DoG). Each grid point is assigned the filter response and points with very negative values outside the protein are defined as cores. Neighbouring cores are merged. Bottlenecks are defined as regions where the intensity gets closer to zero. The dilation of sub-pocket cores by a 2 Å radius towards the protein surface and an assembly of adjacent sub-pockets leads to the complete list of pockets and their constituting sub-pockets. The detected pockets are ranked with respect to the volume of their DoG cores.

McVol[200] Originally, McVol is a Monte Carlo method to estimate molecular volumes by randomly placing points inside a box with the POI and determining whether it is placed inside the molecule or the solvent. The molecular surface is described by probe spheres which probe the hard spheres that represent the molecule. This SAS is represented by points in space. Random points are placed inside the box and their location (inside solvent, protein, or cleft) is assessed. Additionally, a graph-based algorithm is used to find internal cavities and surface clefts. Surface points inside and outside the protein are identified by generating an undirected graph. The surface dots within predefined distances are connected.

Given that the largest subgraph describes the outer protein surface, the smaller subgraphs indicate inner surfaces. A second procedure maps random points which were placed during the Monte Carlo integration on a grid. Points inside an atom, a void, and the solvent are counted. The solvent grid cells that are completely surrounded by protein grid cells are distinguished from the solvent grid cells that are connected to the box borders by another graph-based evaluation. Besides the detection of internal cavities, solvent accessible pockets are identified by evaluating all solvent grid points. The cavity volume is determined by the Monte Carlo integration. No cavity ranking was introduced for the method.

CLIPPERS / Travel Depth[190] Similar to LIGSITE, the protein structure is placed onto a grid and the molecular Connolly surface is calculated for the Travel Depth analyses.[189] The Quickhull algorithm[310] is used to determine the convex hull, i.e., the smallest convex set that contains the surface points of the protein. Subsequently, grid points are assigned as interior, outside the convex hull, and between surfaces. The minimum distance which has to be travelled by a solvent probe to the convex hull as reference is calculated for each surface point. Interior points (defined as cavities) are excluded from further calculations. The advancement of the method offers the inclusion of buried cavities by the addition of a "virtual" edge connecting them to the external surface.[190] The pocket detection tool CLIPPERS (Complete Liberal Inventory of Protein Pockets Elucidating and Reporting on Shape) can be applied which enumerates all regions of the surface within the convex hull. A sorted list with respect to the deepest point of a region is analysed. Pocket points are used to define a new pocket or are added to a new pocket and the pockets are merged. Finally, a filtering step of uniqueness is applied to obtain the set of interesting pockets. Very large (> 2000 Å³) and small (< 25 Å³) pockets are excluded. A ranking of the identified pockets is not provided, but the reported pocket size can be used for this purpose.

CAVE and CAVE-CL[184,185] The method CAVE was developed to detect internal cavities (completely enclosed) in proteins. The generation of an enveloping triangulation[311] enables the assignment of any spatial point towards a cavity based on its relation to the triangulation. This triangulation leads to the tetrahedrons that encircle voids in the protein structures, whereas the tetrahedron's void-intersecting triangles are not considered. The vdW radii of the atoms can be set by the user and they are extended by a probe ball radius which is defined by the radius of a water molecule by default. In 2015, the original version was optimised and supplemented by the OpenCL version of the approach (CAVE-CL). The identified cavities are not ranked, but the volume and the surface area can be used to this end.

fconv[195] Pockets can be identified by a grid-based or a Delaunay triangulation-based approach. The resulting pockets are ranked according to their volume. The volume estimation and robustness with respect to site flexibility is reported to be better for the triangulation-based methodology. For this approach, the lower α level can be decreased to yield finer pockets. The upper α level can be increased to find larger cavities. The defaults are set to 2.5 and 8 Å, respectively. Moreover, cofactors, metal ions, etc. can be provided as additional input to consider them as part of the protein and avoid the identification of already known sites.

MSPocket[202] First, the solvent excluded surface is calculated with the help of the MSMS program[312] resulting in surface vertices which are triangulated. Internal components of the solvent excluded surface are directly reported as "cavities", i.e., interior, solvent inaccessible pockets. For the exterior components (solvent-exposed cavities) the surface triangulation is represented as a graph model. The vertices are labelled with the angle between their normal vector (N_A) and the normal vector of the adjacent vertices. If all angles are smaller than 90°, the average angle is calculated and assigned to the vertex. In contrast, the average angle is set to infinity if one angle exceeds 90°.

The number of vertices is reduced by choosing representative vertices. Cluster vertex pairs are assigned if their distance is less than or equal to 8 Å. Subsequently, pocket vertex pairs and protrusion vertex pairs are defined using cut-off distances ($d_{AB} - d_{A'B'} > 0.2 \cdot 1.3$ Å) for pockets and cut-off angles (sum of the angles between the vector N_A and the segment $\overline{AB}$ and between the vector N_B and the segment $\overline{BA} \geq 200°$) for protrusions. The vertex pairs are translated along their normal by 0.2 Å. The pocket vertices are clustered based on their proximity and their definition. The vertices which are part of a protrusion vertex pair are not considered. An iterative exclusion of the outlier vertices which have not more than one neighbour or not more than four pocket partners enables a pocket pruning. The resulting cavities are ranked according to their volume.

PocketAnalyzer^{PCA}[205] PocketAnalyzer^{PCA} can be applied to process ensembles of protein structures, identify pockets, and analyse their druggability (in combination with SiteMap[219]). Its basic field of application is the analysis of MD simulation-derived structural ensembles. The analysis of pocket shapes and characteristics in structural ensembles enables comprehensive investigations of protein cavities. The grid-based prediction of putative binding sites is similar to LIGSITE[273] and a method developed by Stahl *et al.*[313] However, hydrogen atoms of the proteins are considered. Three criteria must be met by a grid point to be assigned as cavity point. A grid point that does not reside within the vdW radius of any protein atom has to be enclosed by the protein structure. Therefore, its degree of buriedness is calculated by a scan along fourteen vectors (positive and negative directions along the x-, y-, and z-axes and the diagonals) and the calculation of the number of directions where protein atoms are found within a 10 Å distance. The minimum is set to 11. Finally, a minimum number of adjacent buried neighbouring grid points is required (15 by default). The subsequent clustering leads to distinct pockets that are filtered according to cluster size (by default at least 50 for a grid spacing of 0.8 Å). The pocket volume is used for the ranking of the proposed pockets.

CAVITY SEARCH[188] The software was developed at the John Hopkins University, Baltimore, USA. It was designed to identify pockets in proteins by differentiating between two binding site types: cavities and crevices. The predicted sites are not ranked, but their volume is calculated.

Void[212] Void implements a tessellation-based method which is based on the α shape theory.[314] In this approach, the Delaunay triangulation is the basis of all further calculations. The value of α is used to filter the sub-complexes of this triangulation. The detected pockets are ranked according to their surface area.

dxTuber[192] This grid-based method can be applied to identify voids, tunnels, or clefts in proteins. It relies on the input density maps for the protein and solvent which can be generated by VolMap as integrated into VMD.[231] Density voxels can be scanned in three modes. A 3D scanning procedure aims at identifying buried cavities, while the 2D scanning highlights channel-like pockets which are connected to the exterior. The identification of surface clefts is realised via the 1D scanning method.

For the latter procedure, at least one of the three principal axes (x, y, z) has to be chosen. The solvent voxels with a minimum density are analysed for adjacent protein voxels in both directions of the three axes. In the 1D procedure, at least one protein voxel is required in both directions of at least one axis for the solvent voxel to be considered as protein-internal. For 2D and 3D this criterion is extended to at least two and three axes, respectively.

The restrictiveness rises from 1D to 3D, whereas the number of false positives was shown to decrease. A grouping of cavities is realised by walking in the z-, y-, and x-direction and assigning the internal solvent voxels to groups based on their neighbourhood. During this process, groups which were classified separately, but found to be adjacent are merged. For each internal solvent voxel, the 26 neighbouring voxels are analysed with respect to their internal or external character. A threshold for the minimum number of internal neighbouring voxels leads to the exclusion of the voxels on the protein surface. A pocket ranking is not performed, but the number of voxels per detected cavity are reported.

VolSite[50] For the VolSite annotations and site descriptor calculations, a grid for the protein-ligand complex, which is centred on the ligand's centre of mass, is generated. The grid points within the grid cells are annotated according to the distance to any protein atom and the points outside the protein are analysed with respect to their buriedness. Inaccessible voxels are removed using an adjustable threshold. The remaining cavity points are clustered and cavity points with less than three neighbours are also rejected. Pharmacophore properties are assigned to the remaining cavity points in a distance of at least 4 Å of any protein atom, while all other points have a null property. In the unrestricted mode (for pocket prediction), VolSite is applied to the complete protein and the output consists of up to 21 putative binding sites which are sorted by their volume (i.e., the number of pharmacophore-annotated voxels).

CAVITATOR[186] CAVITATOR shows high similarity to LIGSITE, but is said to be less sensitive with respect to structural irregularities. Per definition, the heavy atoms of the surface occupy the central grid point and the neighbouring grid points within √2 Å, i.e., 27 grid points. Points are part of a pocket if they are surrounded by heavy atom occupied grid points in both directions along the x-, y-, or z-axis. The detected pockets are ranked according to their volume.

CAVITY[187] Probe spheres are used for the geometric binding site detection with CAVITY. A 3D grid is generated which includes the complete protein structure. The grid points are assigned as occupied and vacant based on the solvent accessible protein radii. The points between both types are stored as surface grid points. A large sphere with a 10 Å radius is moved across the protein surface and the vacant grid points within the sphere space are removed (outside points). A shrink-and-expand algorithm is applied to further separate the identified cavities. The vacant points connected to any outside point are level-one layer points which are connected to level-two layer points etc. until the bottom layer is reached. A minimal and maximal depth is defined to eliminate or separate cavities. Finally, an empirical CavityScore is calculated for the binding site ranking. It is based on the observation that a large cavity volume, a small entrance area, a large hydrophobic volume, and a small cavity surface area, but a large hydrogen bond forming surface area are reliable indicators for true binding sites.

MOLE 2.0[201] The Delaunay triangulation of all atomic centres is used to construct a Voronoi diagram. This diagram is used as a graph whose vertices correspond to the circumcentres of the Delaunay tetrahedrons. Edges are drawn for two tetrahedrons sharing a common side. The tetrahedrons at the interface between protein and environment are removed if they are large enough to fit a sphere with a given probe radius (large probe) or if they are too small to fit another sphere with a given probe radius (small probe). For the remaining tetrahedrons, which form components based on their connections, cavity diagrams are determined. They include at least one tetrahedron on the molecular surface. The cavities in the output are ranked by their volume.

NanoShaper[203] The NanoShaper procedure starts by calculating the shape of the protein surface. A box is created surrounding the protein. Grid-consistent rays are cast to detect intersections with the surface. These intersections are assembled and the enclosed volume is calculated. This algorithm offers an accurate and robust way for the construction and analysis of the molecular surface. The cavity detection is realised by means of a floodfill algorithm. Volume- and shape-filters are applied to remove unwanted pockets and the surface area of the remaining sites is calculated by a surface triangulation. The finally derived cavities are characterised by their volume and surface area but are not ranked.

BetaVoid[19] The method is based on the initial theory of β-shapes[315] and β-complexes[316] and uses the Lee-Richards SAS[317] to identify voids in proteins. As opposed to the commonly used α-shapes,[318] β-shapes consider the differences between the vdW radii of the underlying protein atoms. The Voronoi diagram of the atoms is constructed and the quasi-triangulation of the diagram is used to construct the β-complex. The aim of the software is void recognition, i.e. cavities within a macromolecule are detected. The void volume is used to rank the identified pockets.

KVFinder[199] The protein structure is placed onto a 3D grid and grid points are assigned to two states: occupied and empty. A small radius probe is rolled over the protein and all overlapping grid points are marked (“Probe In”). Second, the protein surface is sampled by a large radius probe and overlapping points are labelled (“Probe Out”). The difference of the “Probe In” and “Probe Out” grid points is defined as a cavity. The volume and surface area of the predicted cavities are calculated and can be used for ranking purposes.

trj_cavity[210] The protein atoms are projected onto a 3D grid and labelled by their vdW radii. This is followed by a scanning procedure searching for cavities. Cavity voxel neighbours in the x-, y-, and z-directions are iteratively analysed terminating if no more neighbours are found. Three requirements for true cavity voxels have to be fulfilled: they have to be empty and connected to at least one other cavity voxel while being surrounded by the protein in a defined number of directions. The burial state of the cavity corresponds to the number of directions where protein atoms are found, i.e., a completely buried pocket has a burial state of six. Consequently, the number of correctly identified pockets depends on the binding site characteristics (void, invagination, cavity, etc.). The pockets can be ranked by their volume.

GetCleft[197] GetCleft represents a reimplementation of the original SURFNET algorithm and is integrated into the NRGsuite. A crude estimate of cavity volumes enables the ranking of the predicted sites. Additionally, the manual adjustment of the identified binding sites is possible.

PrinCCes[208] The methodology behind PrinCCes (Protein internal Channel & Cavity estimation) is based on a geometric decomposition of protein voids. Voids are identified using a 3D grid and a large as well as a small probe sphere. The algorithm Find Continuous SubSpace is used for the efficient decomposition of the void protein volume. A probe sphere in a void and all accessible volume elements define a common subspace. An elaborate 10 step process leads to the identification of cavities but is also able to detect interior channels. The identified pockets are ranked according to their volume.

Energy-Based Binding Site Prediction Methods

PocketFinder[319] PocketFinder is introduced here because a reimplementation of the algorithm was used for the ConCavity predictions. The protein structure is mapped onto a grid to determine the Lennard-Jones potential (with an aliphatic carbon atom as the probe) at the grid points. A subsequent space averaging of the resulting potential map is used to smooth the obtained map. Putative ligand envelopes are determined by contouring the final map at a predefined level that was empirically determined by applying the method to a training set. In contrast to most other prediction tools, the ligand surface rather than the protein pocket is predicted. Envelopes with a volume below 100 Å are neglected and the remaining ones are ranked according to their volume.

SiteMap[218,219] The protein structures are prepared using Schrödinger routines, i.e., water molecules and crystallization buffer-derived small molecules are removed, the bond orders and formal charges of ligands and cofactors are assigned, hydrogen atoms are added, reorientations of residue side-chains and ligand substituents are performed, and a final structure refinement was applied to relax the modified structures. Subsequently, a 3D grid is built that encompasses all protein atoms. The grid points are classified according to their distance to the surrounding protein atoms and their corresponding vdW radii. A point is considered outside the protein if the ratio of the squares of both distances is above 2.5. The enclosure of each outside point is calculated and the vdW interaction energy with the receptor is determined. Points whose interaction energies are larger than a given threshold and insufficiently enclosed points are rejected. Site point grouping and group merging procedures lead to the predicted sites.

Additionally, hydrophilic, hydrophobic, hydrogen bond donor and acceptor, and metal binding maps are generated based on probe interaction calculations. Finally, the site properties are calculated (number of site points, exposure and enclosure, contact scores, hydrophobic and hydrophilic character, hydrogen bond donor and acceptor character). The SiteScore for the pocket ranking is calculated as the weighted sum of the number of site points, the enclosure score, and the hydrophilic score. Similarly, a druggability score is calculated to distinguish druggable and non-druggable sites. It relies on the same site properties, but the individual weights were adjusted.

Binding Response[14] Binding Response is a docking-based algorithm for the prediction of binding sites. After the protein preparation and surface construction, potential binding regions (concave regions) are assigned with the help of SPHGEN.[320] Spheres are positioned at the surface in a way that they touch two protein surface points and lie on the normal of one of both points. Sphere trimming is performed and the subsequent sphere clustering leads to the selection of the top ten clusters (largest number of spheres). The binding pocket is defined as the volume which is covered by the spheres.

Poses of the test compounds in the binding site are predicted and scored using DOCK4.0.[321] A 1,000 compound database including molecules with a high structural diversity is used. The compounds are characterised by drug-like properties (according to the rules of Lipinski[322] and Oprea[323]). The final Binding Response scores of the analysed binding sites is calculated based on a measure for the binding energy (DOCK interaction energy for ligand atoms within the pre-defined binding site space) and the correct positioning of the ligands within the pre-assigned binding cavity (with respect to the centre of mass of the cavity and the ligand). The overall score is calculated by averaging the scores of all compounds or of a subset of the compounds with the highest scores (e.g., a subset of 900 ligands).

SiteHound[217] SiteHound is a probe interaction-based method. The predicted binding sites are characterised by favourable interactions with either a carbon or a phosphate probe. A regular 3D grid is created and affinity maps (MIFs) on the surface of the protein are generated. This is achieved by applying either AutoGrid[324] for vdW interactions or EasyMifs[217] for electrostatic interactions. A subsequent filtering step eliminates unfavourable interaction energies. The clustering of the remaining points according to the spatial proximity (agglomerative hierarchical clustering) leads to clusters that represent the potential binding sites which are ranked according to the Total Interaction Energy, i.e., the sum of all interaction energies of the cluster members (grid points).

MolSite[215] For this docking-based approach, the protein surface is divided into grid boxes with their centres at the POI's Cα atoms. Only Cα atoms with a minimum distance of 8 Å are used for the subsequent docking with the Sievgene/myPresto[325] software. Three compound libraries were prepared by randomly selecting 10,000 compounds from the LigandBox[326] compound database. Neither the size nor the composition of those sets drastically influences the MolSite overall performance as described in the original publication. The binding site likeness was assessed by three scoring measures: the average value of all compound docking scores, the best docking score, and the sum of the average docking score and the standard deviation. The average score provided the best means to predict ligand binding sites based on the first evaluation of the blind docking approach.

AutoSite[214] This energy-based detection method applies hydrophobic (a carbon atom probe) and hydrogen bond donor and acceptor (hydrogen and oxygen probes) interaction energies to derive the putative binding sites. The exclusion of grid points that show low affinities and the merging of high-affinity points leads to the elucidation of potential ligand atoms and consequently to the identification of putative pockets on the surface of the protein.

Different descriptors per clustered site enable the scoring of the resulting pockets. An empirical score, which is calculated based on the size, the buriedness, and the radius of gyration of the predicted sites, is used for the pocket ranking.

Knowledge-Based Binding Site Prediction Methods

LISE[222] Protein-ligand interaction networks that capture interactions between three protein atoms and two ligand atoms were previously generated and statistically analysed.[327] These investigations led to the identification of typical binding motifs represented by protein atom type triangles. Protein triangles for the POI are derived by choosing atom type triangles with all three sides longer than 2 Å and shorter than 13 Å and at least two sides shorter than 10 Å. The protein triangle of these potential interaction motifs is labelled with a binding site EF derived from the statistical analysis of protein triangles which were found within ligand binding sites and non-ligand binding sites. The triangles with a SAS area above 0 Å² for all atoms and a binding site EF of at least 1 are considered for further analyses.

The triangle scores are calculated based on the binding site EF and the conservation score of the corresponding residues that is obtained from the PSSM of a PSI-BLAST[244] search. To enrich potential binding site residues, a grid representation is used. Each grid point is annotated as empty or occupied. Empty grid points are labelled with the triangle scores of surrounding protein triangles (geometric centre of triangles within 4 Å of the grid point). Spheres with a 5.5 Å radius are generated for each empty grid point and the sum of the scores of all included grid points denotes the final sphere score. The tool is named according to the Ligand Interacting and binding Site Enriched protein triangles. The authors tested the methodology with and without PSSMs and concluded that the conservation scores contributed only to a minor extent to the outcome. Nevertheless, the results could be improved by considering residue conservation.

COFACTOR[220] After the processing of either a sequence (automated comparative modelling with I-TASSER[245]) or a known protein structure, a global similarity search with TM-align[71] is performed. For the most similar templates (i.e., with a TM-score above 0.3), an additional local similarity search is performed. To accomplish this, the MSA for the query sequence is generated with the help of PSI-BLAST[244] and the conserved residues are identified using the Jensen-Shannon divergence.[249] Triplets of the conserved residues and the two lanking ones are reported as binding motifs. Candidate and template site motifs are superimposed, and a binding site similarity score is derived. Based on the highest identified local similarities and the corresponding alignments, the ligand of the template binding site is posed in the predicted site of the POI. Finally, all ligands are clustered to define distinct binding sites based on an 8 Å distance threshold. The additional analysis of the Tanimoto similarity of the ligands within one cluster leads to a further grouping. The final ranking is achieved by means of a confidence score which includes the number of predicted ligands, the local and global similarity to the templates, the sequence identity, and the distance of the predicted ligand towards all other predicted ligands of the cluster. The predicted ligands and binding site residues are provided as the final output.

G-LoSA[221] A local structure comparison of the POI to a non-redundant database of PDB-derived binding site structures based on bound molecules enables the prediction of putative cavities with G-LoSA (Graph-based Local Structure Alignment). To this end, two options are available. First, it is possible to compare the protein structure against the complete binding site library. This procedure is computationally very expensive. The second possibility[243] is the identification of putative binding site residues (on the protein surface) with VisGrid[211] and the comparison of the derived pockets with the binding site library. The score of the matches considers the number of aligned residues and the RMSD of the matching Cα atoms. After assigning the top scored 100 template ligands, an average-linkage clustering is performed with a distance threshold of 3 Å to finally choose the best score of each cluster for ranking. The binding sites are then predicted as the centroid of all ligands in the respective cluster.

In the post-processing step, a grid is centred at the predicted sites and points within 3 Å or in a distance of at least 4.5 Å of any receptor atoms, as well as highly solvent-exposed grid points, are removed. The remaining grid points are clustered with a cut-off of 3.46 Å. Binding sites with less than five grid points are removed leading to a reduced list of predicted pockets. The GA-score, which reflects the chemical feature overlap (hydrogen bond donor and acceptor features, hydroxyl groups, positively and negatively charged atoms, aromatic rings, and aliphatic groups) of the graph-based query-to-template alignment, is finally calculated. The GA-score is also used to rank the putative sites.

P2RANK[223] PRANK[41] was initially published as a tool for pocket prediction rescoring and was shown to improve the performance of fpocket (version 1.0) and ConCavity^{LIGSITE} by rescoring the identified cavities based on an RF (Random Forest) classifier. This classification model was trained on a representative binding site dataset by Chen and colleagues[43] which was previously used for benchmarking purposes. Additionally, it is possible to identify and score pockets with P2RANK.[223] A layer of equally spaced points is generated on the Connolly surface of the POI. Atomic feature vectors are calculated for all exposed atoms and aggregated to generate feature vectors for the Connolly points. They represent a 6 Å spherical neighbourhood. Based on the aforementioned RF classifier, these points obtain scores ranging from 0 (“unligandable”) to 1 (“ligandable”). The underlying features include the protrusion, the temperature factor, the aromaticity, the hydrophilicity, etc. and sum up to vectors consisting of 34 features. The single linkage clustering of all points with a score above 0.35 with a distance cut-off of 3 Å leads to the final binding site predictions. Putative pockets with less than three Connolly points are excluded. Finally, the remaining pockets are ranked. The sum of the scores of all points of one cluster is calculated. The cumulative “ligandability score” for one cluster, i.e., binding site, corresponds to this sum to the power of two. Very small pockets are at the bottom of the ranked lists due to the additive nature of the score.

This scoring procedure is also underlying the rescoring tool PRANK which is currently applicable to rescore ConCavity-, fpocket-, SiteHound-, MetaPocket 2.0-, LISE-, and DeepSite[328]-derived cavities.[329]

Combinations of Different Methodologies

ConCavity[34] ConCavity combines the prediction of three pocket identification methods (LIGSITE,[273] PocketFinder,[319] and SURFNET[209]) with sequence conservation information (Jensen-Shannon divergence[249]) obtained from the sequence-based MSAs. The structure-based methodologies represent reimplementations of the original algorithms. The prediction starts by generating a grid. The grid points are assigned scores based on the prediction methodology. The number of protein-solvent-protein elements is the scoring measure for LIGSITE. The number of spheres overlapping a grid point represents the score for SURFNET. For the energy-based method PocketFinder, the Lennard-Jones potential is used to assign scores to the corresponding grid points. Afterwards, a threshold is determined to assign pockets which fulfil distinct criteria (considering the maximum number of pockets, the total volume of all pockets, the minimum volume for any pocket, the minimum cross-section radius for any pocket, the maximum distance from the protein surface). In contrast to the application of a fixed threshold for pocket pruning, this method avoids that no pocket is detected or that the predicted cavities are too large. After several filtering steps, a mapping from the predicted pockets to the neighbouring residues is performed and the residues in close proximity to the pocket points with high scores are also assigned high scores. This methodology ensures that the final scores reflect the likelihood of the residues of being ligand-binding.

MetaPocket 2.0[171] The software is based on the original MetaPocket[35] approach which combined the predictions of four binding site identification tools to improve the hit rates. MetaPocket 2.0 exploits the predictions of LIGSITE^{CS}, PASS, Q-SiteFinder, SURFNET, fpocket, ghecom, ConCavity^{LIGSITE}, and POCASA to generate meta pocket sites. The three top-ranked pockets of each method are used. Their z-score, which is the standardised ranking metrics for each method (the difference between the score, e.g., the volume, and the mean score divided by the standard deviation), is calculated. The subsequent hierarchical clustering is applied to group all 24 sites (the three highest ranked predictions of each method) according to their spatial overlap with a distance cut-off of 8 Å. The z-score per cluster (sum of the z-scores of the individual binding sites in one cluster) is used to rank the final sites. Pocket residues are mapped using the site points of all predicted sites of one cluster. The surface residues (relative SAS area above 20% according to the NACCESS software[330]) within a 5 Å distance of any grid point are assigned as potential ligand-binding residues.

DEPTH (version 2.0)[224] Each protein residue is assigned its probability of being part of a binding site. Two parameters are used to derive this probability. First, the residue depth is calculated based on the residue's average distance to a bulk solvent molecule after iteratively solvating the protein molecule according to an earlier publication.[172] The minimum number of neighbouring waters that are necessary to define water molecules as non-bulk can be adjusted by the user (4 by default). Second, the SAS area is calculated by applying the approach of Shrake and Rupley.[331] A comparison to these characteristics in known binding sites enables an assignment of ligand-binding regions on the protein surface. Based on the analysis of a training set of 900 protein crystal structures with small molecule ligands, the residue depth and SAS area were calculated for the protein residues of known binding sites.

This enables the determination of probability values for residues of novel proteins by calculating the individual residue depth and solvent accessibility. Five iterations of a PSI-BLAST[244] search against a non-redundant sequence database are performed to compute entropy values from the resulting MSAs. The average of the entropy and probability values is used to estimate the probability of each residue to be a ligand-binding one. A final step includes the removal of residues whose probability falls below a pre-defined threshold (0.5 is recommended for binding sites in multimeric proteins) and the grouping of the cavity residues starting from the residue with the highest probability value.

COACH[36] COACH represents a combined approach that was shown to increase the performance of the individual tools TM-SITE, S-SITE, COFACTOR, FINDSITE,[332] and ConCavity. TM-SITE predictions result from the comparison of the query to known ligand binding sites. To this end, pockets predicted by ConCavity with more than 15 residues and the complete structure are used for a TM-align-based alignment with binding sites of a pre-calculated library of ligand-based sites in the BioLiP[333] database. In contrast, S-SITE is based on sequence profile-profile alignments. The profile for the query is generated using PSI-BLAST.[244] The template profiles are pre-calculated results of PSI-BLAST searches for known structures of the BioLiP[333] library. The Needleman-Wunsch algorithm is applied to identify the most similar template profiles. Additionally, secondary structure information of the protein structures as defined by STRIDE[105] and the predicted secondary structures for the query (PSIPRED[334]) are used to identify suitable template structures.

For both tools, TM-SITE and S-SITE, weighting schemes were introduced to assign the binding residues which are predicted with the highest consensus. FINDSITE[332] applies various threading programs to find template proteins that match the query. COFACTOR and ConCavity were described in more detail above and are applied with default settings. A linear SVM (Support Vector Machine) was introduced for a final prediction of the putative binding site residues. The binding site probability values are calculated for all residues for the five methodologies. The resulting feature vectors are passed to an SVM that was trained on 400 non-redundant proteins. The recent version of the executable is restricted to a combination of COFACTOR, S-SITE, and TM-SITE.

Further Properties of Binding Site Prediction Methods

Irrespective of their basic approaches, prediction methods can be distinguished based on the user requirements. This is summarised below (Figure 4.1.2) for the investigated methods. The applicability as a web server is of interest if the method should be applied only once. High runtimes hamper the application to large datasets. While the predicted sites for binding site comparison purposes should be as accurate as possible, a pocket centre is sufficient for the molecular docking into novel sites. Some methods are restricted to buried and enclosed pockets. Especially docking-based approaches are well suited for ligand-centric predictions, whereas probe-based methods identify binding hot spots for potential ligands. Binding site properties and druggability become of interest for drug design applications. Finally, the specialised analysis of pocket dynamics is possible with some of the presented methods.

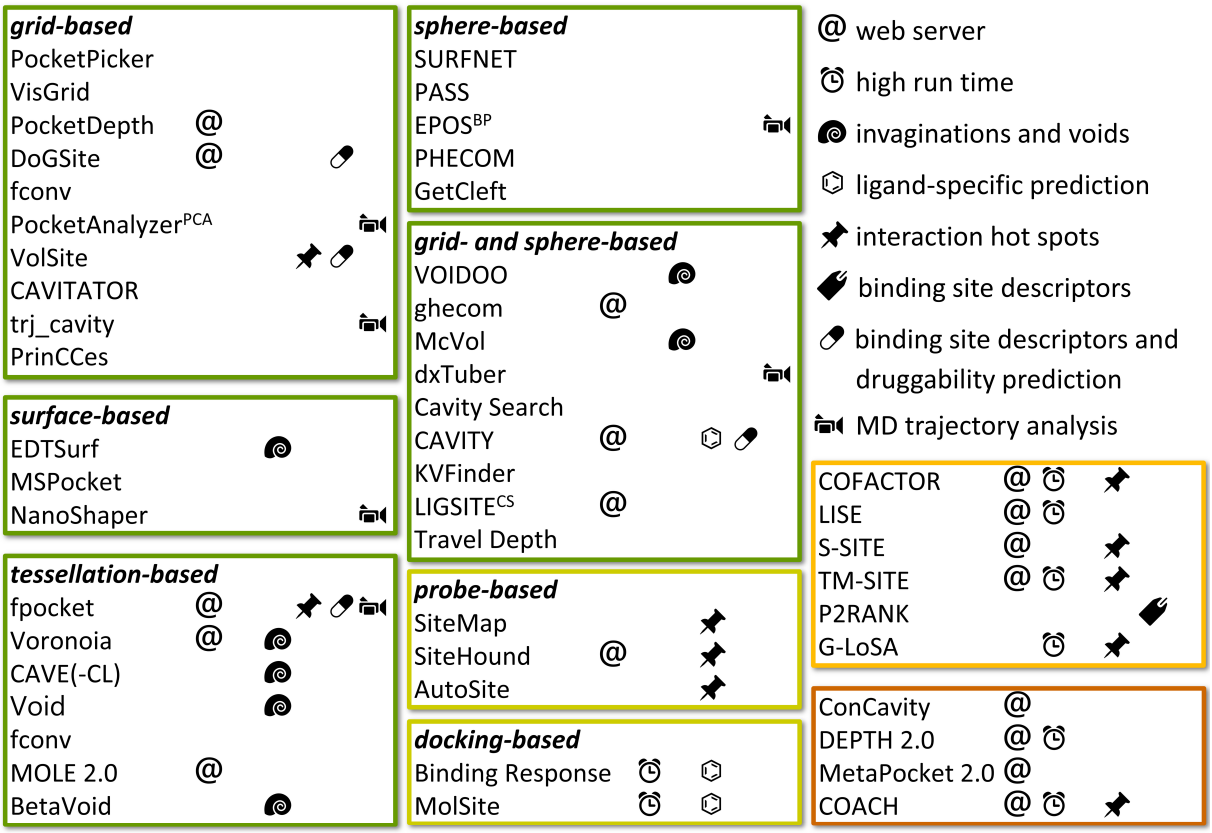


Figure 4.1.2. Special properties of binding site prediction methods. Further criteria which might be relevant depending on the purpose of the predictions are presented for all analysed geometry-based (dark green), energy-based (light green), knowledge-based (yellow), and combined (red) methods.

4.1.3 Identification of Appropriate Benchmark Datasets

In line with the high number of available methodologies to identify and characterise novel binding sites, the number of evaluation datasets is large. In general, the multitude of available structural models enables the generation of datasets for a large-scale evaluation. Nevertheless, such datasets hamper the analysis of time-consuming, but potentially useful tools. Although, there are some datasets that were preferentially used for benchmark analyses (48 bound/unbound,[12] 210 bound structures,[12] 198 drug-target dataset[35]), the majority of the methods was not evaluated by means of a unique state-of-the-art dataset which hampers the choice of an appropriate tool.

Table 4.1.1 provides an overview of the methods analysed herein with respect to the applied evaluation datasets and the methods they were compared to. It highlights the high variability of the benchmark studies that goes along with the challenge of choosing an appropriate tool for the prediction of binding sites. Despite the fact that some objective benchmark analyses were performed,[43] they suffer from the low coverage of existing prediction methods. Intriguingly, the choice of the dataset has a high impact on the ranking of the compared methods, as already shown in the publication of Weisel and co-workers.[207]

Three of the most frequently used datasets (from the MetaPocket 2.0 web site[335]) were downloaded and analysed with respect to the quality of the structures, the biological relevance of the ligands, and the diversity of the dataset.

Table 4.1.1. Evaluation methods for binding site prediction methods. Overview of the tools analysed herein, the benchmark studies for their initial evaluation, and a list of tools they were compared to. Methods in italics represent tools whose aim exceeds the prediction of novel binding sites in protein structure models (e.g., application to MD ensembles, detection of interior cavities).

Method	Benchmark Datasets	Compared to…	Quality Measure
Geometry-Based			
BetaVoid[19]	*100 randomly chosen structures*	*VOIDOO, CASTp*	*exemplary case studies*
CAVE[184,311]	*showcase studies*	-	*detection of inner cavities*
CAVITATOR[186]	-	-	-
CAVITY[187]	setD	LIGSITE^{CSC}, Q-SiteFinder, SURFNET, PASS	overlap of grid points for pockets and ligands
CAVITY SEARCH[188]	-	-	-
CLIPPERS[189,190]	setA	SURFNET, CAST[336]	Tanimoto overlap score of residues
DoGSite[191]	setC, setE	LSite, DSite; VICE, PASS, fpocket, PocketPicker, LIGSITE, CAST, SURFNET	ligand and pocket coverage; geometric centre within 4 Å of any ligand atom
dxTuber[192]	*showcase studies*	*SURFNET, CAVER, PyMOL*	*MD-derived pocket ensemble analysis*
EDTSurf[193]	*six protein cavities*	-	*detection of inner cavities*
EPOS^{BP}[194]	*showcase studies*	-	*prediction of transient binding sites*
fconv[195]	-	-	-
GetCleft[197]	-	-	-
fpocket[196]	setE, filtered setG, setH	PocketPicker, LIGSITE$^{(CS/CSC)}$, CAST, PASS, SURFNET	geometric centre within 4 Å of any ligand atom, overlap of α spheres and ligand atoms
ghecom[198]	dataset of 1,817 chains	Q-SiteFinder, PASS, SURFNET, PHECOM	overlap of grid points for pockets and ligands
KVFinder[199]	setL	MetaPocket 2.0, LIGSITE^{CS}, POCASA, Q-SiteFinder, PASS, SURFNET, ghecom, ConCavity, fpocket	geometric centre within 4 Å of any ligand atom
LIGSITE^{CS}[12]	setE, setF	PASS, SURFNET	geometric centre within 4 Å of any ligand atom
McVol[200]	*showcase studies*	-	*detection of inner cavities, but also solvent accessible clefts*

setA: 67 protein chains from Laskowski *et al.*[38]; **setB:** GOLD set of 75 unique proteins[337]; **setC:** PDBbind generated by Wang *et al.*[338]; **setD:** dataset of Laurie and Jackson[183]; **setE:** set of 48 corresponding holo and apo structures[12]; **setF:** 210 protein structures from the PLD (Protein Ligand Database)[339] extracted by Huang and Schroeder[12]; **setG:** druggability dataset of 63 structures created by Cheng *et al.*[340]; **setH:** Astex diverse set for protein-ligand docking[341]; **setI:** LigASite dataset with 550 structures[342]; **setJ:** dataset of 17,126 structures from An *et al.*[319] as modified by Schmidtke *et al.*[42]; **setK:** 251 ligand-bound structures and 104 similar apo structures from Chen *et al.*[43]; **setL:** a non-redundant set of 198 drug-target structures from DrugPort[343] created by Zhang *et al.*[171]; **setM:** combinations of protein-ligand complexes from different sources[341,342,344]

Table 4.1.1 *(continued)*. Evaluation methods for novel binding site prediction methods. Overview of the tools analysed herein, the benchmark studies for their initial evaluation, and a list of tools they were compared to. Methods in italics represent tools whose aim exceeds the prediction of novel binding sites in protein structure models (e.g., application to MD ensembles, detection of interior cavities).

Method	Benchmark Datasets	Compared to…	Quality Measure
Geometry-Based			
MOLE 2.0[201]	showcase studies	MOLE 1.4, MolAxis, CAVER 2.0, CAVER 3.0	exemplary case studies
MSPocket[202]	setE	VICE, fpocket, POCASA	geometric centre within 4 Å of any ligand atom, overlap between predicted atoms and ligand-contacting atoms
NanoShaper[203]	structure of FAAH	EDTSurf	exemplary case studies
PASS[11]	30 bound and unbound chains	-	exemplary case studies
PHECOM[204]	4,071 chains from SCOP database	PASS, SURFNET	precision ($V_{predicted} \cap V_{ligand}/V_{predicted}$), recall ($V_{predicted} \cap V_{ligand}/V_{ligand}$)
Pocket-Analyzer^{PCA}[205]	*showcase studies (aldose reductase, neuraminidase)*	-	*analysis of MD simulation derived pocket ensembles*
PocketDepth[206]	setB, setC, setD, setE, setF	CASTp, Q-SiteFinder, LIGSITE^{CSC}, LigandFit	true positives show at least 10% volume overlap with the ligand
PocketPicker[207]	setE, setF	CAST, LIGSITE$^{(CS/CSC)}$, PASS, SURFNET	geometric centre within 4 Å to any ligand atom
PrinCCes[208]	showcase studies	-	-
SURFNET[209]	-	-	-
trj_cavity[210]	*showcase studies*	*MDpocket, CAVER 3.0, HOLE*	*analysis of MD simulation derived pocket ensembles*
VOIDOO[20]	-	-	*detection of inner cavities and invaginations*
VolSite[50]	NRDLD[345]	DrugPred, SiteMap, fpocket	analysis of druggability predictions
Voronoia[213]	-	-	-
Energy-Based			
AutoSite[214]	setH	SiteHound, ConCavity (all versions)	Jaccard coefficient ($V_{predicted} \cap V_{ligand}/V_{predicted} \cup V_{ligand}$)
Binding Response[14]	29 complexes from the LPDB[346]	-	percentage of correct top1 pocket predictions

setA: 67 protein chains from Laskowski *et al.*[38]; **setB:** GOLD set of 75 unique proteins[337]; **setC:** PDBbind generated by Wang *et al.*[338]; **setD:** dataset of Laurie and Jackson[183]; **setE:** set of 48 corresponding holo and apo structures[12]; **setF:** 210 protein structures from the PLD[339] extracted by Huang and Schroeder[12]; **setG:** druggability dataset of 63 structures created by Cheng *et al.*[340]; **setH:** Astex diverse set for protein-ligand docking[341]; **setI:** LigASite dataset with 550 structures[342]; **setJ:** dataset of 17,126 structures from An *et al.*[319] as modified by Schmidtke *et al.*[42]; **setK:** 251 ligand-bound structures and 104 similar apo structures from Chen *et al.*[43]; **setL:** a non-redundant set of 198 drug-target structures from DrugPort[343] created by Zhang *et al.*[171]; **setM:** combinations of protein-ligand complexes from different sources[341,342,344]

Table 4.1.1 *(continued)*. Evaluation methods for novel binding site prediction methods. Overview of the tools analysed herein, the benchmark studies for their initial evaluation, and a list of tools they were compared to. Methods in italics represent tools whose aim exceeds the prediction of novel binding sites in protein structure models (e.g., application to MD ensembles, detection of interior cavities).

Method	Benchmark Datasets	Compared to…	Quality Measure
Energy-Based			
MolSite[215]	ligand-bound structures of setE	FINDSITE, SURFNET, MetaPocket, LIGSITE, Q-SiteFinder, PASS	geometric centre within 4 Å to any ligand atom
SiteHound[216,217]	77 protein-ligand complexes	-	distance between the energy-weighted geometric centre of the binding site and the ligand
SiteMap[218,219]	setC	-	geometric centre within 4 Å to any ligand atom
Knowledge-Based			
COFACTOR[220]	setM	FINDSITE, ConCavity	MCC, precision, and recall based on residues
G-LoSA[221]	setM	fpocket, TM-align instead of G-LoSA, CMCS-BSP	least distance between the top5 predicted binding sites and the geometric centre of the ligand (BS error), percentage of sites within a 4 Å BS error
LISE[222]	setE, setF	MetaPocket 2.0, Q-SiteFinder, PASS, LIGSITE$^{CS/CSC}$, VICE, SURFNET, DoGSite, fpocket, MSPocket, POCASA, CAST, PocketPicker	geometric centre within 4 Å to any ligand atom, percentage of grid points within 1.6 Å of any ligand atom[183]
P2RANK[223]	sets E, F, H, L, setJ, setK	fpocket, ConCavity	geometric centre within 4 Å to any ligand atom
S-SITE, TM-SITE[36]	*see COACH*		
Combined Approaches			
COACH[36]	setM	COFACTOR, FINDSITE, ConCavity	MCC, precision, and recall based on residues
ConCavity[34]	setI	LIGSITE$^{(CS/CSC)}$, CASTp, SURFNET, Pocket-Finder, Q-SiteFinder	over-prediction factor ($V_{predicted}/V_{ligand}$), precision, recall, Jaccard coefficient
DEPTH (version 2.0)[224]	filtered setI	ConCavity, MetaPocket 2.0	MCC
MetaPocket 2.0[171]	setE, setF, setL	LIGSITE^{CS}, POCASA, Q-SiteFinder, PASS, SURFNET, ghecom, ConCavity, fpocket	geometric centre within 4 Å to any ligand atom

setA: 67 protein chains from Laskowski *et al.*[38]; **setB:** GOLD set of 75 unique proteins[337]; **setC:** PDBbind generated by Wang *et al.*[338]; **setD:** dataset of Laurie and Jackson[183]; **setE:** set of 48 corresponding holo and apo structures[12]; **setF:** 210 protein structures from the PLD[339] extracted by Huang and Schroeder[12]; **setG:** druggability dataset of 63 structures created by Cheng *et al.*[340]; **setH:** Astex diverse set for protein-ligand docking[341]; **setI:** LigASite dataset with 550 structures[342]; **setJ:** dataset of 17,126 structures from An *et al.*[319] as modified by Schmidtke *et al.*[42]; **setK:** 251 ligand-bound structures and 104 similar apo structures from Chen *et al.*[43]; **setL:** a non-redundant set of 198 drug-target structures from DrugPort[343] created by Zhang *et al.*[171]; **setM:** combinations of protein-ligand complexes from different sources[341,342,344]

The 48 bound/unbound dataset was excluded due to the low number of structures and the fact that some of the X-ray structures had a high resolution or R-factors above 0.25. Structures of intermediate quality should be avoided to ensure that the tools do not miss binding sites due to problems with the structural model. Analysing the dataset of 210 bound structures, one also comes across structures of insufficient quality. Additionally, the dataset includes NMR structures whose quality is difficult to assess.[347] For other structures, it is not self-evident which ligand binding site should be considered. For some complexes, cofactors are included in the binding site definition. Furthermore, the size of the ligands varies from small drug-like molecules to large polysaccharides. The dataset was derived from the former manually curated database PLD v1.3[339] and redundant proteins were removed. However, structures without conservation scores in the ConSurf-HSSP database were excluded leading to a biased dataset as the sequence and functional diversity of the protein structures cannot be guaranteed. Finally, the protein-ligand complex structures in the 198 drug-target dataset were investigated. Although this dataset comprises known drug-target binding sites, it includes redundant protein structures in varying numbers for different protein classes. An exclusion of duplicate entries led to 170 structures. Nevertheless, an evaluation of the biological significance of the ligands defining the binding sites of interest led to further problems besides the already mentioned ones. Buffer components such as DMSO or guanidine were selected as binding site-defining ligands. Although their location might hint at an accessible site, it cannot be guaranteed that this site is a biological relevant one.

Finally, the “gold-standard dataset” of the biologically relevant binding sites in protein structures, LigASite,[342] was considered. The non-redundant version of the dataset comprises 391 structures, but no unique HETCODES are provided for a reliable re-processing of the pockets. In contrast to the previous datasets, cofactors and metal ions are an integral part of the ligands in the dataset. They are often involved in ternary binding sites. Moreover, they are of minor interest in the identification of druggable sites and often detectable via protein and binding site comparison or metal binding site prediction.

These observations led to the decision to create a novel dataset using diverse X-ray structures of high quality in complex with drug-like molecules. The biological significance of the analysed ligands was ensured by manual curation. Aside from the challenges discussed in the previous paragraphs, the binding affinity must be considered. Low-affinity complexes are outside the aim of the present study. Moreover, some of the analysed tools might fail for non-standard protein residues in the proximity of the binding site.

Finally, an independent dataset was generated based on the PDBbind core set.[225] It fulfils all of the quality requirements mentioned above (structure quality, no missing atoms, no modified residues within the immediate proximity of the ligand, no ternary complexes, at least 15% buried ligand surface area). Moreover, the set includes three protein-ligand complexes per protein class, which enables a better approximation of the binding site dimensions. Additional apo structures corresponding to the respective protein-ligand complexes were included to consider different conformational states of the binding sites.

The visual inspection of the matches between apo and ligand-bound structures and an analysis of the RMSD of the resulting sequence-based structure alignments (see Table A1.3) showed insignificant binding site changes upon ligand binding. As the methods' performance for a more challenging set of binding sites is of even more relevance, the dataset of cryptic site pairs of Cimermancic and co-workers[18] was additionally adjusted and applied for the evaluation of the different binding site prediction methodologies.

Two further datasets were prepared to assess the method's applicability to challenging binding sites. First, the problem of predicting allosteric sites was addressed. Second, it was investigated whether inaccuracies in theoretical protein structure models are tolerated.

4.1.4 Challenges for Binding Site Prediction Benchmark Studies

As a general prerequisite for the evaluation of binding site prediction methods, some special aspects of binding site identification must be considered. Depending on the protein class and function, the number of small molecule binding sites can vary. In addition to catalytic sites, regulatory pockets and varying numbers of sub-pockets can be found. The analysis of the knowledge-based small molecule binding site database IBIS (Inferred Biomolecular Interaction Server) led to the finding that the average number of biologically relevant binding sites per chain is approximately four.[348] This is usually incorporated in prediction benchmark studies by calculating hit rates for the top-ranked (top1) and the three highest ranked pockets (top3).

Moreover, the size of the ligand does not necessarily correspond to the binding site volume. The problem of defining the exact binding site dimensions can never be completely solved. Ligand-based definitions are biased due to the presence of sub-pockets which are not addressed or the binding of multiple ligands in one pocket. Additionally, binding site flexibility hampers a unique definition of a small molecule binding site. The definition of the pocket of interest varies from one protein-ligand complex to another. This is illustrated in Figure 4.1.3 for a small molecule binding site in four structures of human coagulation factor Xa. This problem was partially circumvented by using "composite ligands", i.e., the ligands of all three complex structures of every protein class in the extended PDBbind core set[225] were used to define the small molecule binding site to minimise the ligand bias.

4.1.5 Quality Criteria for the Evaluation of Binding Site Prediction Methodologies

The high number of available methods shows the manifoldness of requirements for a suitable binding site prediction tool. The available methods were evaluated based on several quality criteria that might be of interest to a varying extent depending on the scope. First, the time for the search for putative pockets becomes a bottleneck for several approaches. Computationally expensive methods such as blind docking or huge database searches will become infeasible when aiming at the prediction of binding sites for hundreds of structures, MD snapshots, or even the whole PDB. In contrast, they are applicable to the functional annotation of a single protein structure, the analysis of allosteric pockets in an enzyme, or the characterisation of multiple binding sites of one structure, e.g., to elucidate druggable binding sites.

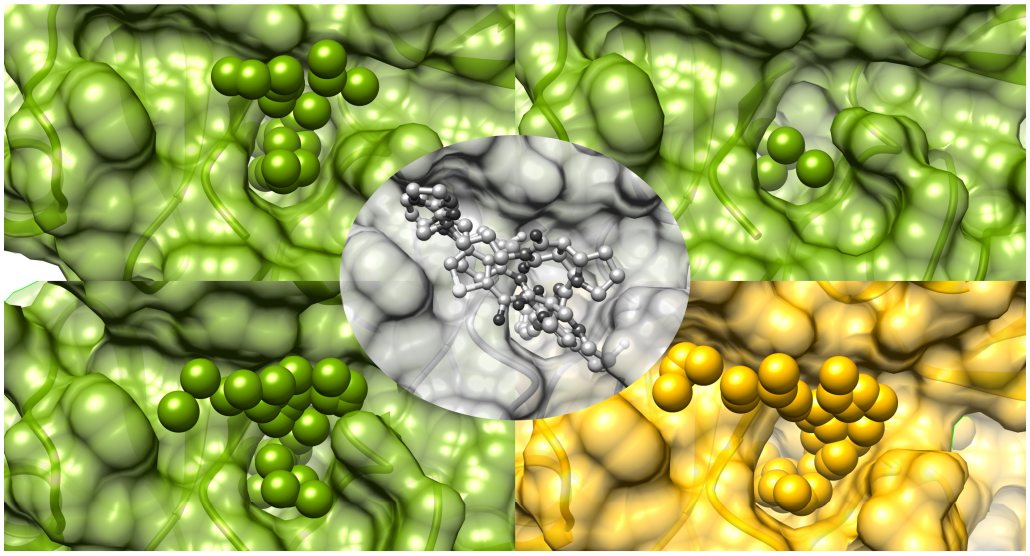
The methods' ability to pinpoint pockets varies from reporting a single point in space to a set of points (grid) to the potentially ligand-contacting residues. Moreover, the question arises whether the methods can identify pockets not only in ligand-occupied structures, but also in apo structures of proteins. To investigate this in more detail, ligand-free structures are included in two of the datasets. Finally, the challenge of so-called cryptic binding sites was addressed. The common approach for correctly predicting such pockets is the generation of protein structure ensembles via MD simulations and their subsequent geometric analysis. Nevertheless, similarity-based methods might be able to correctly assign cavity residues as binding hot-spots even without the necessity of elaborate MD simulations which require the proper structure preparation, computational resources, and the researcher's expertise.

The quality criteria used in this study were chosen by considering the special challenges of large-scale evaluations. First, the output of the tools differs to a huge extent. While some tools (LIGSITE^{CS}, LISE, PASS, SURFNET) report binding sites by a pocket centre with its corresponding coordinates, other tools provide grids or even the binding site-lining residues. A residue-based evaluation (*RRO*, see Equation 1.1 in Section 3.1) favours tools reporting putative binding site residues. For pocket grids, residues in different pre-defined radii can be used for the performance assessment as they map the cavity dimensions. However, for pocket centres, this evaluation will lead to low hit rates as both, solvent accessible and buried residues, will be classified as binding ones for non-spherical sites.

Therefore, the second quality criterion (*RLO*, see Equation 1.5 in Section 3.1) assesses the tools' performance based on pocket centres. They are easily derived by calculating the arithmetic means of all reported pocket grid point coordinates or reported protein atoms. The percentage of ligand atoms within a pre-defined distance of the pocket centre is calculated to analyse the performance. This measure enables the reliable evaluation of methods whose output is restricted to pocket centres. Additionally, it provides information on the accuracy of the methods.

Figure 4.1.3 illustrates the meaning of both quality measures for the structures of coagulation factor Xa from the extended PDBbind core dataset. In this case, all predictions hint at the correct small molecule binding site. Based on the bound ligands of the three holo structures, the prediction for the structure with the PDB-ID 1hcg is the best one. This prediction is ranked much higher with respect to the *RRO*, but shows a comparable *RLO* with respect to the prediction for the other structures. Consequently, the *RRO* provides an estimate of the accuracy of the predicted pocket, but its meaning is biased by the varying cavity definitions. The detected cavity of the structure with the PDB-ID 1mq6 would not be regarded as correct although it gives a crude estimate of the ligand binding site. In contrast, the *RLO* provides a means to evaluate such matches independent of the shape and dimensions of the predicted pocket. The determination of hit rates necessitates a definition of thresholds for the quality measures which range from 0 to 1. These thresholds were varied to estimate the accuracy of the best performing tools within this study.

PDB-ID: 2y5h (*RRO* = 0.52, *RLO* = 0.99) **PDB-ID: 1mq6** (*RRO* = 0.48, *RLO* = 0.95)



PDB-ID: 2xbv (*RRO* = 0.72, *RLO* = 1.00) **PDB-ID: 1hcg** (*RRO* = 0.75, *RLO* = 1.00)

Figure 4.1.3. P2RANK binding site predictions for X-ray structures of human coagulation factor Xa. The identified pockets are represented as grid points and the quality measures for the performance assessment (*RRO*, relative residue overlap, and *RLO*, relative ligand overlap) are given in parentheses. The apo structure is highlighted in yellow. The grey insert shows the ligands of all three ligand-bound structures in the structure with the PDB-ID 1hcg.

4.1.6 The Extended PDBbind Core Set

The first type of evaluation for the PDBbind core dataset is based on the *RRO* of the methods to investigate the residue-based accuracy of the binding site predictions (Figure 4.1.4). At first glance, the overall low hit rates meet the eye. None of the tools shows a hit rate above 80%, i.e., every fifth pocket is not correctly identified and ranked within the top3 pockets. In general, we observe a superiority of energy- and knowledge-based methods over purely geometry-based ones. AutoSite (energy-based) and P2RANK (knowledge-based) outperform the remaining tools if only the highest ranked pocket is considered. The combined method ConCavity^{SURFNET} (conservation) is the best performing tool with respect to the top3 ranked sites. This trend is observed for the complete dataset as well as the ligand-free protein structures. The hit rate decreases by 8% if residue conservation information from HSSP files is not used (ConCavity^{SURFNET}). In contrast, the inclusion of conservation scores only slightly improves the performance of ConCavity using the reimplemented LIGSITE and PocketFinder methodology. Intriguingly, ConCavity^{LIGSITE} was originally recommended for the identification of pockets with ConCavity.[34]

The application of time-consuming tools does not necessarily lead to a notable rise in the hit rates. Of these methods, Binding Response and LISE led to the overall highest top3 hit rates. Even, meta approaches do not result in higher hit rates compared to the usage of single tools. Although MetaPocket 2.0 predictions do not consider POCASA and Q-SiteFinder results as in the originally published approach, the combination of six methods does not improve the hit rates. The hit rates are even lower as compared to ConCavity^{LIGSITE} alone which is one of the six combined methods suggesting the huge impact of POCASA and Q-SiteFinder prediction on the overall performance of MetaPocket 2.0.

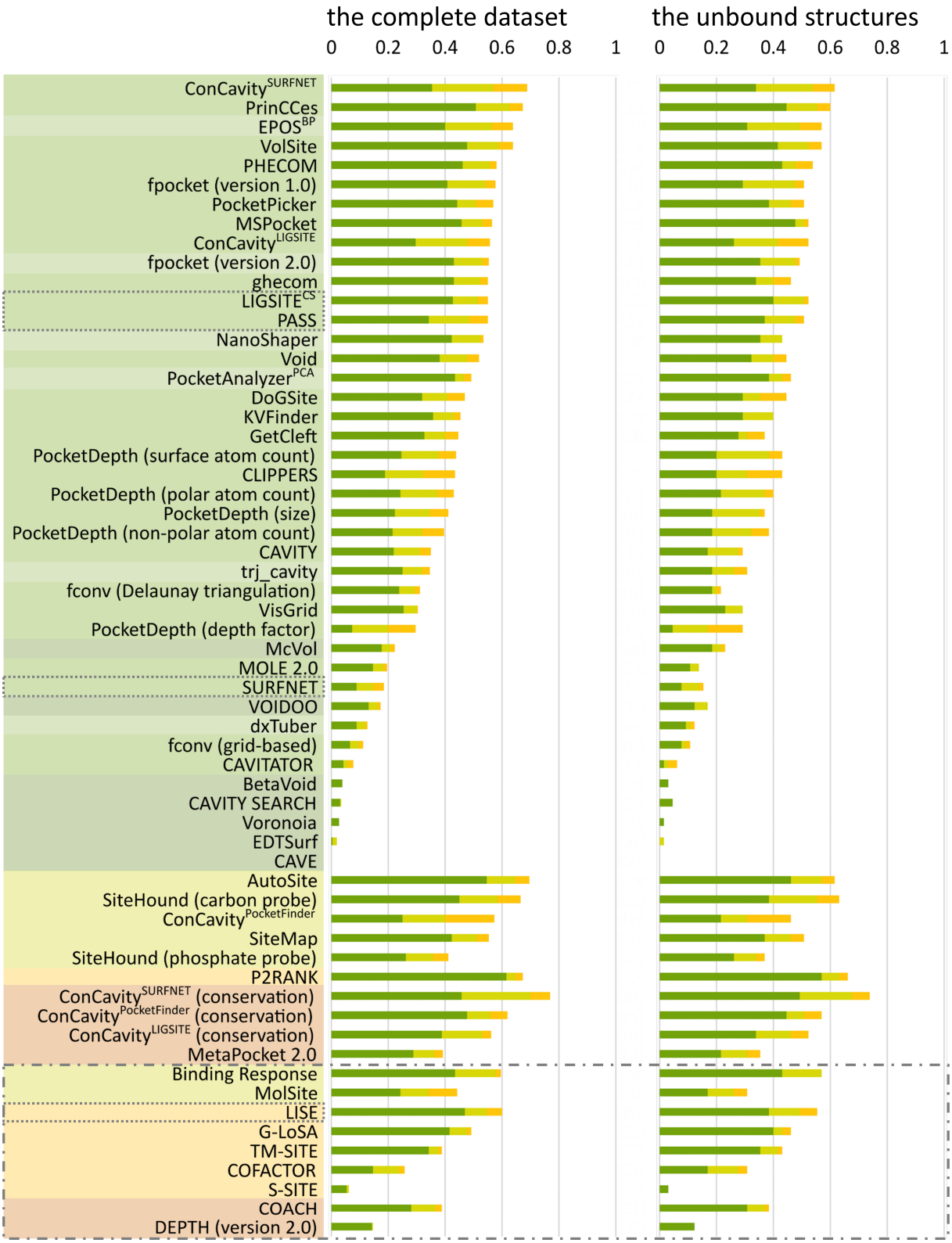


Figure 4.1.4. Residue-based evaluation of prediction tools for the extended PDBbind core set. The top1 (dark green), top2 (light green), and top3 (yellow) pocket hit rates (*RRO* ≥ 0.5) of different binding site identification methodologies for the extended PDBbind core dataset are shown. The chart on the left gives the hit rates for the complete dataset, while the chart on the right represents the hit rates for the 65 apo structures. Methods highlighted in dark green, light green, yellow, and red predict binding sites based on geometry, energy, knowledge, and a combination thereof, respectively. The box (dot-dashed) highlights tools which are computationally very expensive. Dashed boxes highlight methods which annotate predicted sites by a single point. The light and dark grey boxes are used for tools that can be applied to MD trajectories and to identify interior protein binding sites, respectively.

The *RRO*-based hit rates of G-LoSA in combination with VisGrid-identified pockets (recommended way for binding site prediction) can be increased by nearly 5% if the complete protein structures are compared to the binding sites in the library. Intriguingly, the hit rates for the unbound structures do not increase significantly (the *RRO*-based hit rates change from 0.4, 0.43, and 0.46 to 0.42, 0.42, and 0.48 for the top1, top2, and top3 ranked pockets, respectively). The overall increased hit rate results from a better performance for the holo structures. Here, the *RRO*-based hit rates rise from 0.42, 0.49, and 0.50 to 0.49, 0.54, and 0.57 for the top1, top2, and top3 ranked pockets, respectively. Consequently, the tool ranks within the 15 best for the extended PDBbind core set if the complete protein is used for the binding site comparison. Nevertheless, the runtime increases by a factor of nearly four for this dataset and G-LoSA seems to be very sensitive towards the binding site state (apo/holo).

PocketDepth provides five metrics for the ranking of the predicted pockets. While the surface atom count, the binding site size, and the count of non-polar or polar atoms lead to the best results for this dataset, the depth factor is obviously unsuitable for the identification of the most relevant sites of the extended PDBbind core dataset. This holds especially true with regard to the top1 hit rates and the trend persists for the subsequently analysed datasets.

Comparing the performance for the complete dataset and the dataset without the holo structures shows that minor conformational variances in the structures are well tolerated by most tools. The highest mean RMSD values between the apo and holo structures are observed for the antibody FAB fragments (PDB-ID 4isv, PDB-ID 1q9k) and glutamate receptor 2 (PDB-ID 4u2r). The RMSD values are 3.99, 2.44, and 2.27 Å, respectively. The maximum decrease in the hit rates was observed for MolSite (hit $rate_{all}$ - hit $rate_{apo}$ = 0.13 for the top3 pockets). However, this trend can be expected for a docking-based tool which relies on reasonable binding site conformations for the molecular docking.

However, this loss in performance for apo sites cannot be observed for the second docking-based method Binding Response (hit $rate_{all}$ - hit $rate_{apo}$ = 0.03 for the top3 pockets). One basic difference between both methods is the way, the putative binding sites for docking are assigned. While MolSite uses a distinct set of evenly clustered protein Cα atoms as docking centres, Binding Response predicts pockets using SPHGEN which results in a more accurate definition of the site dimensions. Additionally, the compound library applied for Binding Response was generated with a special focus on drug-like compound properties and diversity, whereas the MolSite libraries were randomly generated. In contrast to the docking procedure of Binding Response, which starts with a ligand breaking step and the identification of potential binding site anchors for the largest molecule fragment (DOCK4.0[321]), MolSite applies the Sievgene/myPresto software.[325] The prediction process of eight steps starts with the pocket initiation, the calculation of the electrostatic potential on the SAS of the site, and a ligand conformer generation. Subsequently, a global search is performed to find preferable poses of the ligand conformers on the receptor surface. This is the step which will probably fail if the binding site conformation in the apo state is significantly different from the one in the ligand-bound state. This might be the reason for the sensitivity towards binding site states.

Very low hit rates (below 20%) are observed for SURFNET, VOIDOO, dxTuber, the grid-based algorithm implemented in fconv, CAVITATOR, BetaVoid, CAVITY SEARCH, Voronoia, EDTSurf, CAVE, S-SITE, and DEPTH (version 2.0). It must be noted that BetaVoid, Voronoia, EDTSurf, and CAVE were developed for the investigation of interior protein pockets which are completely enclosed by the protein. This type of pocket is underrepresented in the PDBbind core set. Moreover, VOIDOO and McVol are well suited for the detection of inner pockets, but the detection of clefts depends on the pocket depth and enclosure. These trends are reflected by the correlation between the hit rates and geometric pocket properties as calculated by DoGSite (Table 4.1.2). While a high surface area to volume ratio (enclosed and small binding sites) often results in high hit rates for the top3 cavities, a high lid area correlates negatively with the corresponding hit rates for all aforementioned tools except VOIDOO. Remarkably, a similar trend is also observed for PocketAnalyzer^{PCA} and MOLE 2.0. Most of the remaining tools are characterised by a poor performance for proteins with a high surface to volume ratio. Consequently, the low hit rates of BetaVoid, Voronoia, EDTSurf, CAVE, and VOIDOO result from the small number of buried and interior binding sites within the dataset. Despite the low hit rates for the five tools, it might be beneficial to include them for the detection of interior or buried cavities as other tools might fail to detect them.

Another interesting trend in Table 4.1.2 is observed for LISE, but also to a minor extent for fpocket (version 1.0), CAVITY, and ConCavity in combination with conservation data. They perform better for less hydrophobic pockets with many aromatic and hydrogen bond donor/acceptor atoms. This might hint at a bias in their pocket scoring procedure. The scores are usually derived from pocket datasets. A tendency toward distinct binding site types in those datasets can lead to such trends. For the ConCavity counterparts that do not exploit residue conservation data, this trend diminishes. This finding suggests an additional bias resulting from the sequence conservation data. However, for other methods that rank pockets based on scores rather than by geometric properties (volume, area), this trend is not that pronounced. The methods fpocket (version 2.0), SiteMap, AutoSite, P2RANK, G-LoSA, TM-SITE, COFACTOR, S-SITE, COACH, DEPTH, and MetaPocket 2.0 do not show this correlation and should be preferentially used to identify binding sites independent of their physicochemical nature. Overall, the top3 hit rates of CAVE, P2RANk, and COACH were least influenced by any site descriptor analysed herein.

Next, the influence of the *RRO* threshold, which is used to distinguish between correctly predicted and unidentified pockets, was analysed with respect to the hit rates. The ranking of identification methods for an *RRO* of at least 0.5 changes dramatically when using a different threshold. Depending on the needs of the user, a higher *RRO* threshold might be informative for the search for the most appropriate tool. While a docking into an identified site might not suffer from a coarse binding site definition, the comparison of known ligand-binding cavities to predicted sites might be influenced by the residues which are regarded to be the binding ones. The performance of all tools for four more thresholds (*RRO* ≥ 0.6, 0.7, 0.8, and 0.9) was therefore investigated and the results are presented in Figure 4.1.5.

Table 4.1.2. Pearson correlation coefficients of the methods' top3 hit rates (*RRO* ≥ 0.5) and geometric and physicochemical binding site descriptors calculated with DoGSite. The number of donor, acceptor and aromatic atoms was normalised by the number of site atoms prior to the calculation of correlation coefficients. A colour gradient from red (-0.5) to yellow (0) to green (0.5) was applied to highlight trends.

Method	Lid	Depth	Surface/Volume	Lid/Hull	Hydrophobicity	Acceptor Atom Fraction	Donor Atom Fraction	Aromatic Atom Fraction
ConCavity^{SURFNET}	0.07	0.10	-0.17	-0.12	-0.13	0.04	0.12	0.12
PrinCCes	0.05	0.14	-0.19	-0.24	0.08	-0.18	0.03	0.03
EPOS^{BP}	0.16	0.11	-0.31	0.02	-0.17	0.02	0.20	0.20
VolSite	0.13	0.18	-0.23	-0.19	0.03	-0.15	0.05	0.05
PHECOM	0.07	0.15	-0.22	-0.11	-0.07	-0.09	0.11	0.11
fpocket (version 1.0)	0.08	0.17	-0.07	-0.15	-0.26	0.14	0.23	0.23
PocketPicker	0.04	0.06	-0.23	-0.03	-0.12	-0.06	0.17	0.17
MSPocket	0.00	-0.09	-0.24	0.00	-0.03	-0.03	0.03	0.03
ConCavity^{LIGSITE}	0.13	0.20	-0.13	-0.23	-0.04	-0.07	0.12	0.12
fpocket (version 2.0)	0.06	0.14	-0.10	-0.24	0.01	-0.14	0.09	0.09
ghecom	0.08	0.06	-0.31	-0.02	-0.11	-0.09	0.17	0.17
LIGSITE^{CS}	0.09	0.10	-0.19	-0.07	-0.19	-0.01	0.26	0.26
PASS	0.20	0.24	-0.18	-0.13	-0.17	0.07	0.17	0.17
NanoShaper	0.04	0.06	-0.19	-0.18	0.10	-0.16	-0.08	-0.08
Void	0.04	0.08	-0.24	-0.16	0.08	-0.16	0.01	0.01
PocketAnalyzer^{PCA}	-0.12	0.07	0.10	-0.46	0.12	-0.21	0.01	0.01
DoGSite	-0.07	-0.18	-0.36	0.12	-0.09	-0.07	0.18	0.18
KVFinder	-0.04	0.01	-0.15	-0.24	0.01	-0.13	0.12	0.12
GetCleft	0.04	0.03	-0.29	0.03	0.00	-0.07	0.08	0.08
PocketDepth (surface atoms)	0.20	0.14	-0.29	0.06	-0.08	-0.01	0.10	0.10
CLIPPERS	0.09	0.01	-0.15	0.02	-0.14	-0.01	0.24	0.24
PocketDepth (polar atoms)	0.14	0.07	-0.26	0.10	-0.15	0.07	0.13	0.13
PocketDepth (size)	0.17	0.02	-0.32	0.15	-0.02	-0.03	0.02	0.02
PocketDepth (non-polar atoms)	0.20	0.11	-0.29	0.12	-0.05	-0.07	0.09	0.09
CAVITY	-0.14	-0.14	-0.02	-0.10	-0.21	0.02	0.24	0.24
trj_cavity	0.04	0.01	-0.25	-0.04	-0.09	-0.17	0.25	0.25
fconv (Delaunay triangulation)	0.04	0.09	-0.13	-0.16	0.11	-0.28	0.09	0.09
VisGrid	-0.06	0.14	0.09	-0.32	0.05	-0.16	0.03	0.03
PocketDepth (depth factor)	0.14	0.08	-0.23	0.05	-0.06	-0.04	0.06	0.06
McVol	-0.22	0.01	0.34	-0.47	0.02	-0.09	0.09	0.09
MOLE 2.0	-0.05	0.19	0.15	-0.43	0.07	-0.07	-0.03	-0.03
SURFNET	0.19	0.09	-0.25	0.11	-0.11	0.00	0.15	0.15
VOIDOO	0.00	0.31	0.10	-0.35	0.05	-0.04	0.01	0.01
dxTuber	0.20	0.01	-0.38	0.17	0.01	-0.18	0.11	0.11
fconv (grid-based)	-0.02	-0.13	-0.37	0.10	0.04	-0.23	0.12	0.12
CAVITATOR	0.20	0.02	-0.33	0.20	-0.07	-0.04	0.10	0.10
BetaVoid	-0.14	0.15	0.32	-0.30	-0.13	0.07	0.20	0.20
CAVITY SEARCH	-0.10	0.20	0.33	-0.35	0.02	0.01	0.04	0.04
Voronoia	-0.12	0.13	0.31	-0.32	-0.07	0.02	0.15	0.15
EDTSurf	-0.08	0.06	0.25	-0.25	0.02	0.02	0.03	0.03
CAVE	0.10	0.11	-0.06	0.01	0.02	0.05	-0.05	-0.05

Table 4.1.2 *(continued)*. Pearson correlation coefficients of the methods' top3 hit rates (*RRO* ≥ 0.5) and geometric and physicochemical binding site descriptors calculated with DoGSite. The number of donor, acceptor and aromatic atoms was normalised by the number of site atoms prior to the calculation of correlation coefficients. A colour gradient from red (-0.5) to yellow (0) to green (0.5) was applied to highlight trends.

Method	Lid	Depth	Surface/Volume	Lid/Hull	Hydrophobicity	Acceptor Atom Fraction	Donor Atom Fraction	Aromatic Atom Fraction
AutoSite	0.07	0.13	-0.18	-0.17	-0.12	-0.02	0.19	0.19
SiteHound (carbon probe)	-0.01	0.08	-0.14	-0.17	-0.15	0.09	0.09	0.09
ConCavity^{PocketFinder}	0.25	0.08	-0.35	0.13	-0.14	0.03	0.12	0.12
SiteMap	-0.01	0.06	-0.17	-0.13	-0.17	-0.03	0.28	0.28
SiteHound (phosphate probe)	0.08	0.12	-0.11	-0.18	-0.01	-0.13	0.23	0.23
P2RANK	-0.06	-0.10	-0.05	-0.03	-0.09	0.07	0.08	0.08
ConCavity^{SURFNET} (conservation)	-0.01	0.01	0.11	-0.18	-0.24	0.10	0.27	0.27
ConCavity^{PocketFinder} (conservation)	0.18	0.14	-0.19	0.00	-0.30	0.10	0.30	0.30
ConCavity^{LIGSITE} (conservation)	0.19	0.23	-0.11	-0.19	-0.06	-0.12	0.22	0.22
MetaPocket 2.0	0.02	-0.02	-0.23	-0.05	-0.13	-0.09	0.22	0.22
Binding Response	0.13	0.03	-0.18	0.01	-0.14	0.05	0.09	0.09
MolSite	0.02	0.01	-0.24	-0.09	-0.09	-0.03	0.13	0.13
LISE	-0.14	-0.01	0.23	-0.22	-0.37	0.21	0.44	0.44
G-LoSA	0.11	0.16	-0.01	-0.21	0.11	-0.16	-0.03	-0.03
TM-SITE	0.20	0.12	0.01	-0.03	0.02	-0.02	0.01	0.01
COFACTOR	0.09	-0.04	-0.15	0.07	0.03	-0.04	0.01	0.01
S-SITE	0.12	0.03	-0.14	0.12	0.01	0.01	-0.02	-0.02
COACH	0.16	0.08	0.00	-0.04	-0.06	0.07	0.03	0.03
DEPTH (version 2.0)	0.20	0.16	-0.21	-0.09	-0.02	-0.08	0.10	0.10

In general, a higher *RRO* threshold results in lower hit rates for the top3 predicted pockets. This drop in the hit rates is most pronounced for the pocekt centre-reporting methods LIGSITE^{CS}, LISE, and PASS underlining the difficulty of reliably assessing their performance based on the residue overlap. Additionally, CAVITY and SiteHound (using a carbon probe) hit rates decrease by nearly 25% upon increasing the *RRO* threshold from 0.5 to 0.6. As already shown for the *RRO* threshold of 0.5 (see Figure 4.1.4), the use of SiteHound with a carbon probe leads to considerably higher hit rates than the use of SiteHound in combination with a phosphate probe.

Regardless of the threshold, PrinCCes and ConCavity^{SURFNET} (with and without conservation data) lead to the highest hit rates. Additionally, the most promising energy-based method AutoSite provides very accurate binding site predictions. These methods should be regarded as the most appropriate ones if reliable and exact binding site definitions were necessary.

Overall, Figure 4.1.5 hints at the challenging issue of detecting correct binding site boundaries and defining crucial ligand-binding residues. At an *RRO* threshold of 0.9, the hit rate of none of the tools exceeds 10%. The best performing tool for this threshold (PrinCCes) shows a hit rate of only 6.5%.

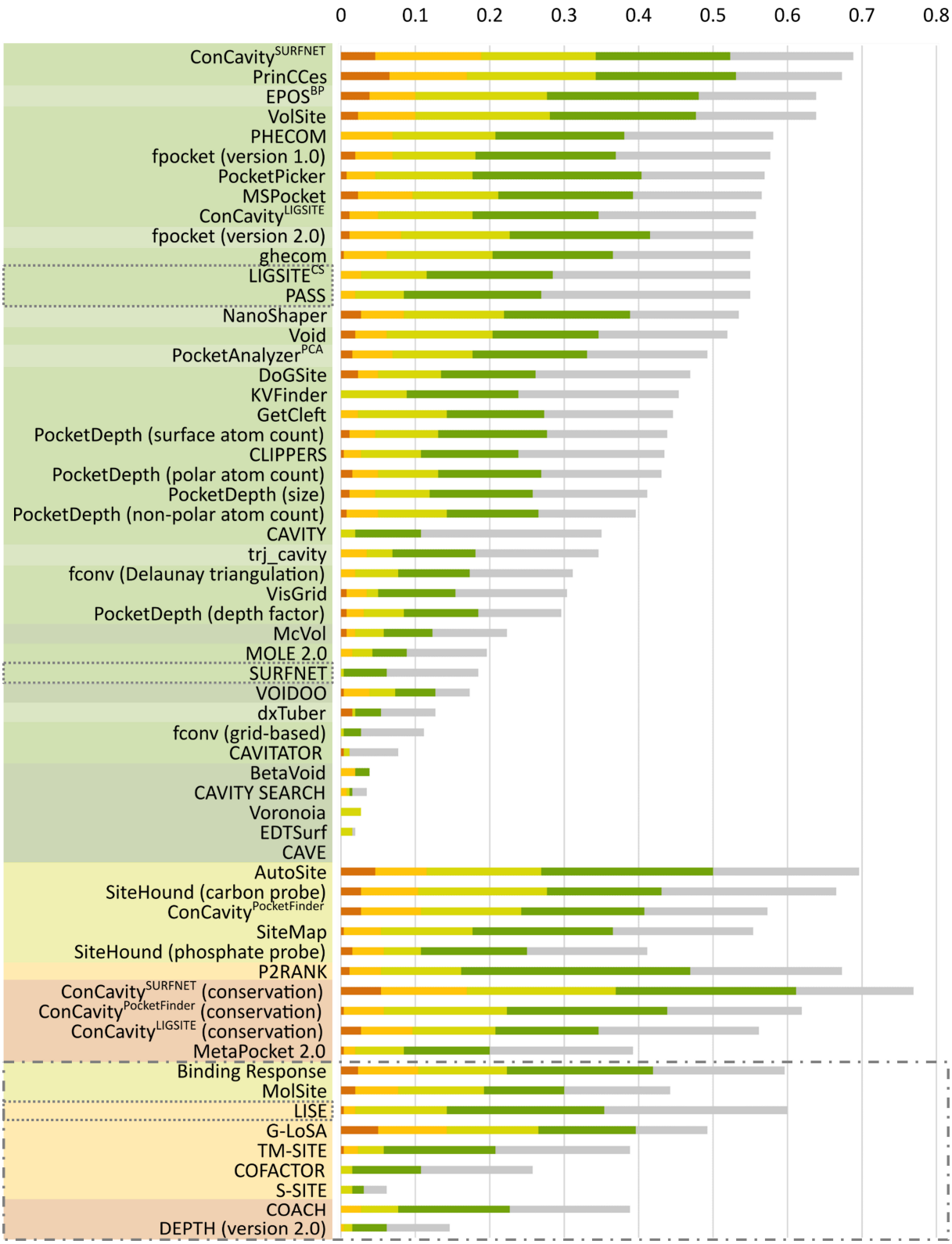


Figure 4.1.5. Threshold dependence of the residue-based evaluation of prediction tools for the extended PDBbind core set. The top3 pocket hit rates for an *RRO* of at least 0.5 (grey), 0.6 (dark green), 0.7 (light green), 0.8 (yellow), and 0.9 (red) of different binding site identification methodologies are shown. Methods highlighted in dark green, light green, yellow, and red predict binding sites based on geometry, energy, knowledge, and a combination thereof, respectively. The box (dot-dashed) highlights tools which are computationally very expensive. Dashed boxes highlight methods which annotate predicted sites by a single point. The light and dark grey boxes are used for tools that can be applied to MD trajectories and to detect interior protein binding sites, respectively.

The residue-based evaluation is well suited for all tools that provide either residues or pocket grids as a detailed pocket representation. In contrast, LIGSITE^{CS}, PASS, SURFNET, and LISE report only the pocket centres. The evaluation of correctly identified residues based on all residues within a certain radius of the respective points is suited for spherically shaped pockets, but might falsify the hit rates for ellipsoid or branched sites. Consequently, the *RLO* measure was used to calculate the hit rates for an *RLO* threshold of 0.5. The results of this evaluation are given in Figure 4.1.6.

The hit rates are higher for all methods if compared to the *RRO*-based hit rates. The ranking of the 60 analysed methods based on the top3 hit rates for the residue-based evaluation (*RRO*) led to the assignment of ranks 26, 29, 37, and 49 for LISE, LIGSITE^{CS}, PASS, and SURFNET, respectively. After a reassessment of the hit rates using the *RLO*, the ranks of LISE and LIGSITE^{CS} increased significantly to 6 and 8, respectively. In contrast, PASS and SURFNET are found at ranks 30 and 45. It can be concluded that both methods, LISE and LIGSITE^{CS}, accurately predict the site location. The residue-based analysis is not suited to examine the performance of these tools. In contrast, the poor performance of PASS and SURFNET prevails in the ligand-based evaluation hinting at difficulties in the prediction of the correct sites.

For SURNET, this finding is most interesting. Its reimplementation in the software ConCavity (even in the absence of conservation data) led to some of the most promising results. An explanation might be the more elaborate pocket extraction (i.e., the search for pockets with reasonable shapes) and residue mapping in ConCavity. The sites which were correctly predicted by ConCavity^{SURFNET} (within the top3 pockets), but missed by SURFNET, were chosen to further investigate this hypothesis (Figure 4.1.7). SURFNET did not correctly identify the binding site of urokinase-type plasminogen activator, whereas ConCavity^{SURFNET} assigned the correct site. The top3 cavities are predicted at other locations on the protein surface. A visualisation of all sites predicted with SURFNET even shows that none of the points is found in the proximity of the ligand. The site of α-L-fucosidase was correctly predicted by ConCavity^{SURFNET}. For the SURFNET prediction, none of the identified sites is situated at the site of interest in the apo structure. Finally, the results for the dimer of HIV-1 integrase were investigated. Both methods fail to predict the correct binding site for the apo structure. For the ligand-bound structures, ConCavity^{SURFNET} assigns the correct site, while none of the SURFNET-derived centres hints at the pocket of interest.

All SURFNET predictions are found outside the ligand-bound cavity. This observation is in line with the low Pearson correlation coefficient of 0.28 for the maximum achieved *RRO* in the top3 pockets for all binding sites predicted by ConCavity^{SURFNET} and SURFNET. Consequently, not only the pocket mapping but also the pocket detection procedure is different. The SURFNET algorithm assigns grid points the value 1 or 0 if spheres overlap this grid point or not, respectively. In the ConCavity reimplementation, grid points that have overlapping spheres are assigned the number of spheres to identify the pocket centre. This difference and the additional filtering and pocket extraction steps in ConCavity probably account for the much higher hit rates of ConCavity^{SURFNET} when compared to those of SURFNET.

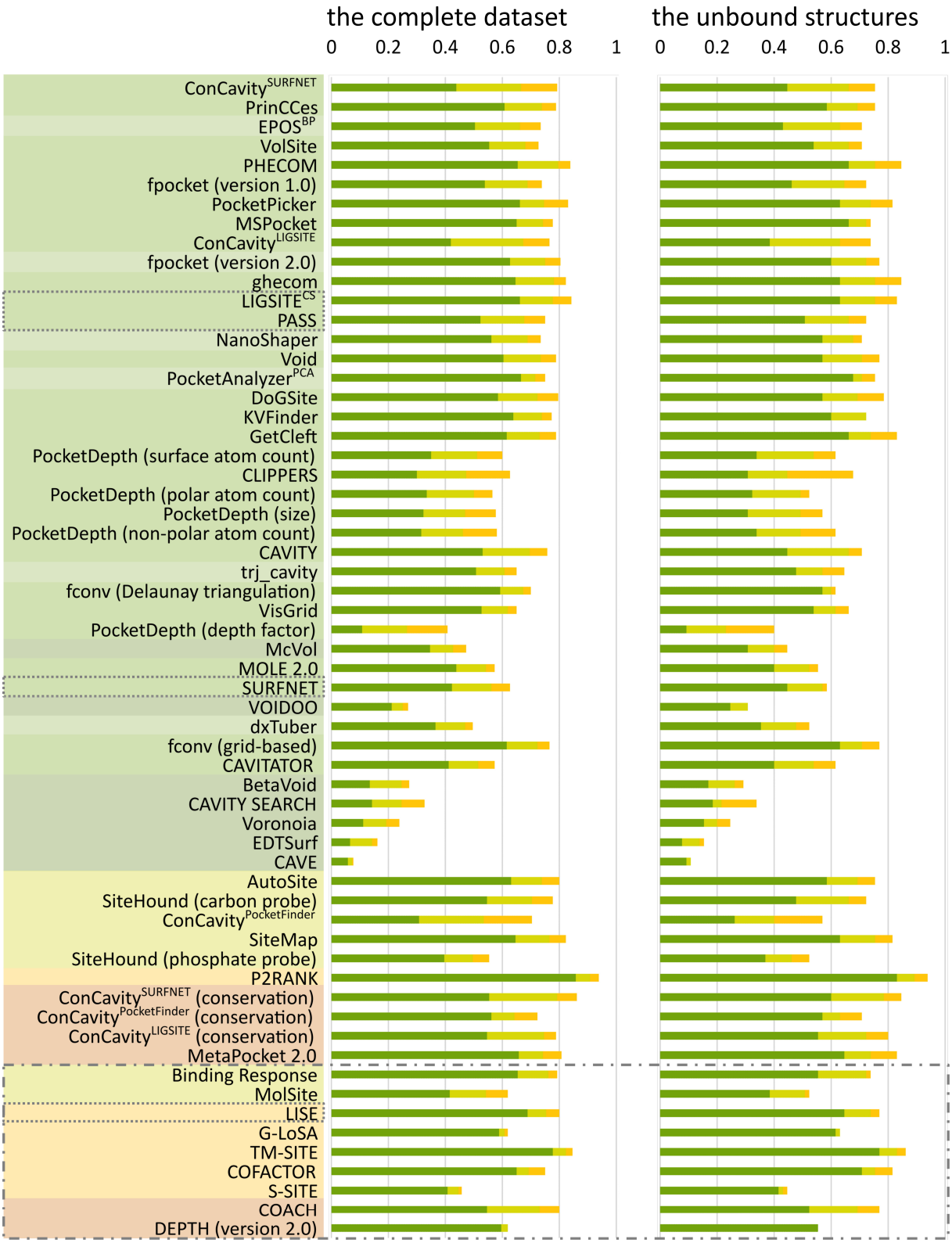


Figure 4.1.6. Ligand-based evaluation of prediction tools for the extended PDBbind core set. The top1 (dark green), top2 (light green), and top3 (yellow) pocket hit rates (*RLO* ≥ 0.5) of different binding site identification methodologies are given. The chart on the left gives the hit rates for the complete dataset, while the chart on the right represents the hit rates for the 65 apo structures. Methods highlighted in dark green, light green, yellow, and red predict binding sites based on geometry, energy, knowledge, and a combination thereof, respectively. The box (dot-dashed) highlights tools which are computationally very expensive. Dashed boxes highlight methods which annotate predicted sites by a single point. The light and dark grey boxes are used for tools that can be applied to MD trajectories and to detect interior protein binding sites, respectively.

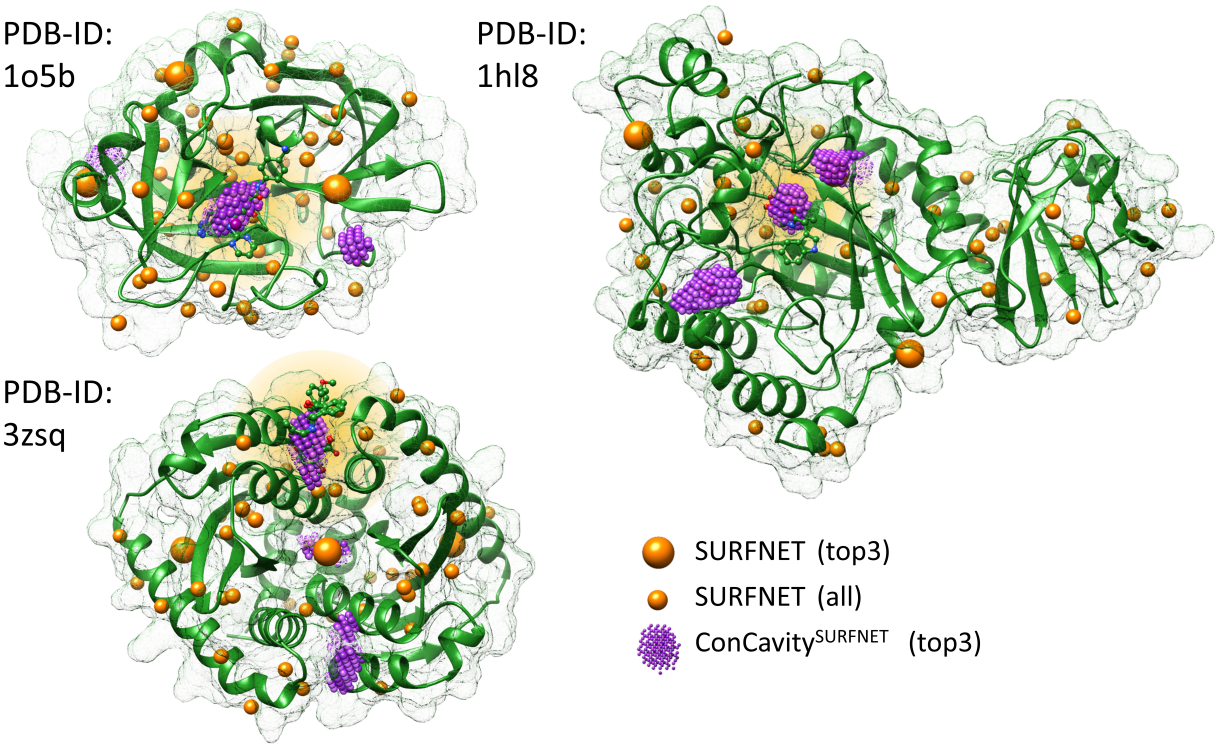


Figure 4.1.7. Differences between SURFNET and its reimplementation ConCavity^{SURFNET}. Binding site predictions of SURFNET (orange spheres) and ConCavity^{SURFNET} (purple spheres) for urokinase-type plasminogen activator (PDB-ID 1o5b), α-L-fucosidase (PDB-ID 1hl8), and HIV-1 integrase (PDB-ID 3zsq) are presented to highlight distinct differences in the predictions of both methods. All presented sites (orange background and ligand in ball-and-stick representation) were correctly predicted within the top3 pockets with ConCavity^{SURFNET}, while the SURFNET algorithm failed to detect these sites.

The results for all binding sites (maximum *RRO* within the top3 pockets) for ConCavity^{LIGSITE} and LIGSITE^{CS} show a higher Pearson correlation coefficient of 0.47. Both methods have similar hit rates in the residue-based evaluation. Nevertheless, differences are observed as the original LIGSITE algorithm was used for the reimplementation in ConCavity. Intriguingly, LIGSITE^{CS} ranks the binding sites of interest higher resulting in a much higher hit rate with respect to the top1 predicted pockets.

Considerable improvements in the hit rates of ConCavity^{LIGSITE} were achieved using the pocket rescoring methodology PRANK in an earlier study.[41] PRANK was initially tested for a rescoring of pockets predicted by fpocket (version 2.0) and ConCavity. Up to 24% and 16% improvement in the hit rates was reported for fpocket and ConCavity, respectively. A decrease in hit rates was never observed. Therefore, all predictions of fpocket and the six versions of ConCavity were rescored using PRANK to evaluate the impact of this rescoring procedure for the dataset (Table 4.1.3). The fpocket results improve with respect to the top1 and top3 hit rates after rescoring for both quality metrics applied in this study (*RRO* and *RLO*). In contrast, only slightly increased or even decreased hit rates are observed for ConCavity. Especially ConCavity^{SURFNET} (with and without conservation data) results worsen significantly for the top1 pockets, while the top3 hit rates did not change. None of the tools performed significantly better than P2RANK before and after rescoring. The only exception is the slightly higher residue-based performance of fpocket (version 1.0). Intriguingly, a significant improvement of the ConCavity results after the rescoring with PRANK could not be observed.

Table 4.1.3. Pocket rescoring with PRANK for the extended PDBbind core set. The top3 and top1 (in parentheses) hit rates (based on *RRO* and *RLO*) of the pocket predictions for the extended PDBbind core set before and after rescoring with PRANK are shown. The hit rate for P2RANK is additionally given for comparison purposes. The difference between the hit rate after rescoring and the initial hit rate is given (Δ) and highlighted in bold characters if a hit rate increase by more than 5% was observed.

Method	Hit Rate for an *RRO* ≥ 0.5			Hit Rate for an *RLO* ≥ 0.5		
	Original	***After Rescoring***	***Δ***	***Original***	***After Rescoring***	***Δ***
fpocket (version 1.0)	0.577 (0.408)	0.677 (0.546)	**0.100** (**0.139**)	0.738 (0.538)	0.862 (0.731)	**0.123** (**0.192**)
fpocket (version 2.0)	0.554 (0.431)	0.596 (0.496)	0.042 (**0.065**)	0.804 (0.627)	0.858 (0.754)	**0.054** (**0.127**)
ConCavity^{LIGSITE}	0.558 (0.296)	0.535 (0.323)	-0.023 (0.027)	0.765 (0.419)	0.738 (0.473)	-0.027 (**0.054**)
ConCavity^{LIGSITE} (conservation)	0.562 (0.388)	0.554 (0.377)	-0.008 (-0.012)	0.788 (0.546)	0.796 (0.542)	0.008 (-0.004)
ConCavity^{PocketFinder}	0.573 (0.250)	0.542 (0.292)	-0.031 (0.042)	0.704 (0.308)	0.662 (0.350)	-0.042 (0.042)
ConCavity^{PocketFinder} (conservation)	0.619 (0.477)	0.627 (0.523)	0.008 (0.046)	0.723 (0.562)	0.746 (0.619)	0.023 (**0.058**)
ConCavity^{SURFNET}	0.688 (0.354)	0.688 (0.223)	0.000 (-0.131)	0.792 (0.438)	0.792 (0.288)	0.000 (-0.150)
ConCavity^{SURFNET} (conservation)	0.769 (0.458)	0.769 (0.396)	0.000 (-0.062)	0.862 (0.554)	0.862 (0.485)	0.000 (-0.069)
P2RANK	0.673 (0.615)			0.938 (0.868)		

Given that the PDBbind core set includes four structures and consequently four identically located sites per protein, the general robustness of the methods with respect to the flexibility of the binding sites can be assessed. To this end, the structures were grouped according to the corresponding protein. The number of hits per group and method was analysed, i.e., each method could achieve a maximum of four correctly identified binding sites per protein. A consistency of 1 was assigned per group and method if none or all of the binding sites were correctly identified. For one or three correctly identified sites, the consistency of 0.5 was assigned, while the consistency was set to 0 if only two of four sites were correctly predicted. Finally, the mean consistency and standard deviation for each tool were calculated for all 65 groups of structures. The results are given for the best performing tools in Table 4.1.4.

ConCavity^{SURFNET} (conservation) and P2RANK show the highest consistency which was independent of the quality measure (*RRO*/*RLO*). This indicates their robustness with respect to small conformational variations. The consistency of computationally more expensive tools is also worth analysing for this dataset. On a residue-based and ligand-based level, the consistency of TM-SITE is extraordinarily high. This knowledge-based method outperforms all other tools with comparable performance. On a purely ligand-based level, the performance of LISE is similar. The residue-based consistency is lower as the LISE output is restricted to less accurate pocket centres. The low consistency of Binding Response can be partially attributed to the sensitivity of docking-based approaches with respect to minor conformational variations.

Accordingly, the high consistency of the knowledge-based methods (TM-SITE, P2RANK, LISE, COFACTOR, COACH) is easily explicable. If a similar template structure is found or the method was trained on similar structures, the prediction will work for all four structures in one group. In general, the most predictive methods consistently detect the sites within this dataset.

Table 4.1.4. Prediction consistency for the extended PDBbind core set. The mean (Ø) consistency and standard deviation (±) for the best performing (hit rate ≥ 0.8 for *RLO* ≥ 0.5) fast binding site prediction tools for the extended PDBbind core set are shown. The mean consistency and standard deviation of the best performing (hit rate > 0.75 for *RLO* ≥ 0.5) computationally expensive methods are given in italics.

RRO			***RLO***		
Consistency of…	**Ø**	**±**	**Consistency of…**	**Ø**	**±**
TM-SITE	*0.931*	*0.173*	P2RANK	0.954	0.145
ConCavity^{SURFNET} (conservation)	0.877	0.264	*TM-SITE*	*0.954*	*0.169*
P2RANK	0.854	0.259	*LISE*	*0.954*	*0.145*
COACH	*0.838*	*0.305*	ConCavity^{SURFNET} (conservation)	0.923	0.180
COFACTOR	*0.838*	*0.305*	SiteMap	0.923	0.201
LISE	*0.800*	*0.326*	*COACH*	*0.908*	*0.246*
PHECOM	0.777	0.340	LIGSITE^{CS}	0.900	0.251
LIGSITE^{CS}	0.777	0.329	PHECOM	0.892	0.270
AutoSite	0.762	0.363	fpocket (version 2.0)	0.885	0.274
ghecom	0.746	0.341	ghecom	0.877	0.264
fpocket (version 2.0)	0.723	0.317	AutoSite	0.862	0.270
MetaPocket 2.0	0.708	0.360	MetaPocket 2.0	0.862	0.284
PocketPicker	0.708	0.360	PocketPicker	0.862	0.298
SiteMap	0.708	0.360	*COFACTOR*	*0.838*	*0.279*
Binding Response	*0.700*	*0.349*	*Binding Response*	*0.831*	*0.307*

Finally, “difficult” or “cryptic” sites in this dataset were investigated, i.e., binding sites which were correctly identified by only a small fraction of all tested tools. Only four of all 260 sites of the dataset were correctly predicted by at most five of all 60 methods (using the criterion of an *RLO* of at least 0.5). This number increases to 13 if the computationally expensive (mostly knowledge-based) methods are excluded from the analysis indicating their impact on binding site prediction for demanding scenarios.

The characteristics of the most challenging sites can be derived. Some examples are given in Figure 4.1.8. Most tools failed to identify the binding sites of the androgen receptor and transthyretin. For the androgen receptor structures, no tool was able to correctly identify the comparatively shallow and small binding pocket which was characterised as an allosteric and regulatory site of the receptor. The tool dxTuber (which showed an overall mediocre performance) correctly identified all transthyretin sites which are also very flat. A bad hit rate for nearly all tools was observed for different sugar-binding and sugar-degrading enzymes (EC 3.2.1, glycosidases). The deep and buried binding sites of the enzymes α-mannosidase II, β-mannosidase, maltase-glucoamylase, and β-galactosidase were identified by at most 34% of all tools. TM-SITE was the only tool which correctly identified all sites in glycosidases underlining the importance of binding site and structure comparison for reliable binding site predictions. Most of the geometry-based tools fail for those structures due to the buried character and small size of the cavities. Nevertheless, TM-SITE failed to detect the binding sites of fucose-binding lectin PA-IIL.

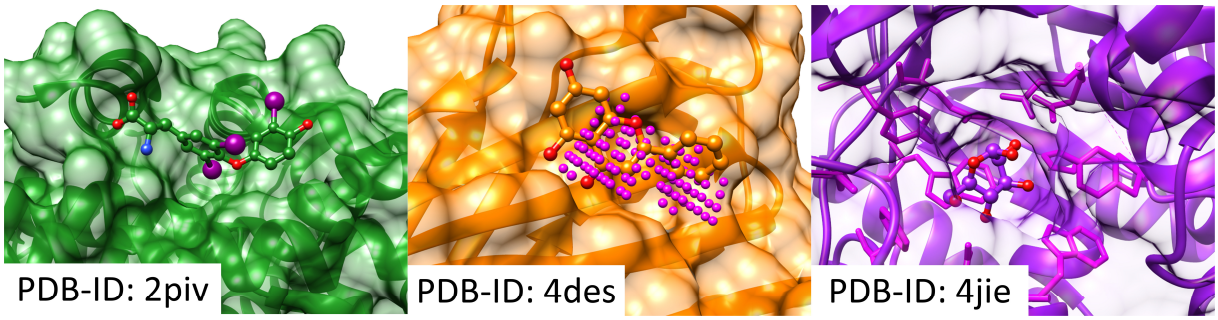


Figure 4.1.8. Challenging sites in the extended PDBbind core set. None of the tools herein correctly predicted the regulatory site of the androgen receptor (green, PDB-ID 2piv). The transthyretin binding site (orange, PDB-ID 4des) was correctly identified by only five tools. The best performing tool was dxTuber (the predicted pocket grid points are given by magenta spheres). TM-SITE and only six further tools predicted the sugar-binding site of β-mannosidase/β-glucosidase (purple, PDB-ID 4jie). The predicted binding residues of TM-SITE are shown as magenta sticks.

4.1.7 The Dataset of Cryptic Binding Sites

The last examples of challenging cases suggested some basic pitfalls for the prediction of small molecule binding sites. Cimermancic and co-workers[18] designed a dataset focusing on the cavities which could neither be detected by fpocket nor by ConCavity. The authors extracted known binding sites from the PDB and identified proteins which are also available in their unbound forms. Residues with at least one atom within 5 Å of any ligand atom were used as the binding site residues. After the removal of redundant and similar (sequence identity > 40%) structures, the remaining protein pairs were analysed with fpocket and ConCavity. Cryptic sites were defined as the sites that were assigned high pocket scores in the bound, but low pocket scores in the unbound conformation. The final original dataset of cryptic sites contains 96 pairs of cryptic sites in the ligand-bound and unbound state.

The complete dataset was considered and all site pairs were superimposed to analyse potential reasons for the failure of fpocket and ConCavity to detect the binding sites. For five pairs, the overall alignment showed a significantly different overall fold of the protein structures in the apo and the ligand-bound state (Figure 4.1.9, left panel). It was assumed that nearly none of the tools will be able to identify the unbound sites without previous analyses of the protein dynamics. Therefore, these pairs were excluded from these analyses. Additionally, the extraction of ligand-binding residues in the proximity of the ligand is precarious and would be biased for these examples.

A reliable ligand-based evaluation of the binding site predictions for the apo structures is hard to achieve as it relies on the ligand positioning after the alignment of the structure pairs to define a common site. Tools which report pocket grids or centres rather than residues (nearly 30% of all prediction tools analysed herein) will be discriminated against on account of a residue-based evaluation. These considerations led to the exclusion of six more pocket pairs as the ligand atoms completely overlap with the protein atoms in the apo state of the protein (Figure 4.1.9, right panel). Consequently, the final set applied herein comprises only 81 apo-holo pairs used for the benchmark of all prediction methods.

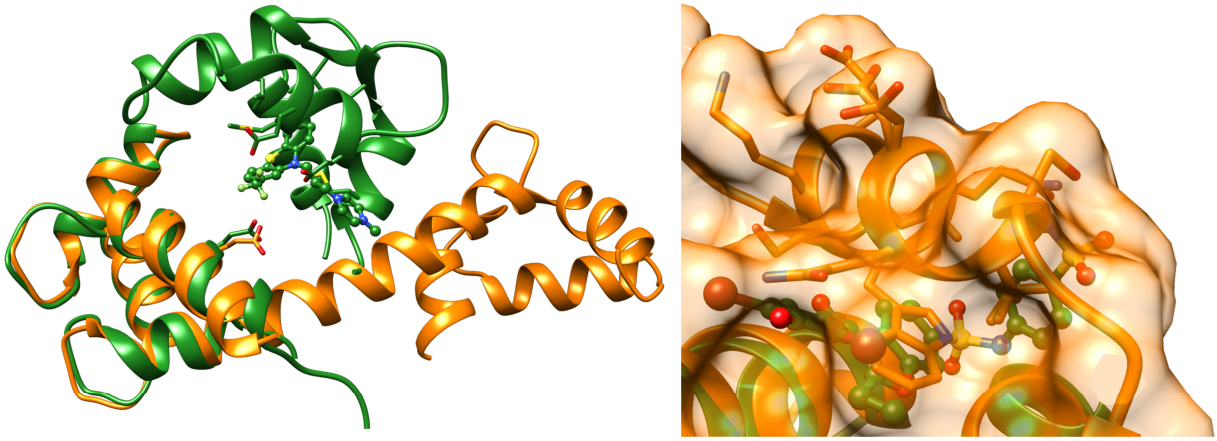


Figure 4.1.9. Examples of cryptic sites excluded from the dataset of cryptic sites before evaluating the binding site prediction methods. Left: Alignment of the structures of calmodulin in the ligand-bound state (PDB-ID 1ctr, green) and the unbound state (PDB-ID 1cll, orange). Residues of both structures in a 4 Å radius of the ligand with the ID TFP are represented as sticks. Right: Alignment of the ligand-bound (PDB-ID 1t49, forest green) and unbound (PDB-ID 2f6v, orange) structures of protein tyrosine phosphatase 1B. The ligand with the ID 892 (ball-and-stick, green) is completely enclosed by protein residue atoms (orange sticks) in the unbound state as indicated by the orange surface representation. A helix is located at the binding site in the apo state.

The RMSD between the protein pairs ranges from 0.24 to 3.74 Å and 63% of all pairs show an RMSD above 1 Å. For comparison, only 20% of the apo structures had a mean RMSD above 1 Å to the holo structures in the previously analysed extended PDBbind core set. Moreover, the RMSD values for the binding sites (as calculated with ProBiS) range from 0.16 to 1.59 Å.

This range is similar to that for the mean PDBbind core RMSD values (0.16 to 1.69 Å). Nevertheless, a larger number of structure pairs with high RMSD values is observed (Figure 4.1.10 (A)). Moreover, this analysis is biased towards binding site elements found to be similar with ProBiS. The flip of only one residue side-chain blocks or partially fills the binding site, but the RMSD to the ligand-bound structure with a different side-chain orientation does not reflect this difference (Figure 4.1.10 (B)). The RMSD is also not influenced by loop movements as the underlying residues are not matched (Figure 4.1.10 (C)).

Large structural variations correlate with the huge differences between the hit rates (*RRO* ≥ 0.5) for the complete set and the apo sites (Figure 4.1.11). Two factors influence these results. First, we saw that the tools' ability to correctly identify sites partially depends on the properties of the cavities. So, there will be sites which are not identified, independent of the state (bound/unbound). Second, conformational deviations between the apo and holo states lead to a decreased ability of the tools to detect these sites. Consequently, it is of interest to identify methods whose performance is least sensitive towards such changes. NanoShaper, SiteHound, PocketDepth (using the depth factor or non-polar atom count), PrinCCes, fpocket, EPOS^{BP} and ConCavity^{SURFNET} showed a much worse *RRO* for at least 40% of the ligand-free structures (RRO_{holo} - RRO_{apo} > 0.3). The same is to be expected and holds true for the docking-based methods MolSite and Binding Response. For some tools, we observe a low number of apo structures whose sets are predicted worse than those of the holo structures. This can be attributed to their overall low performance.

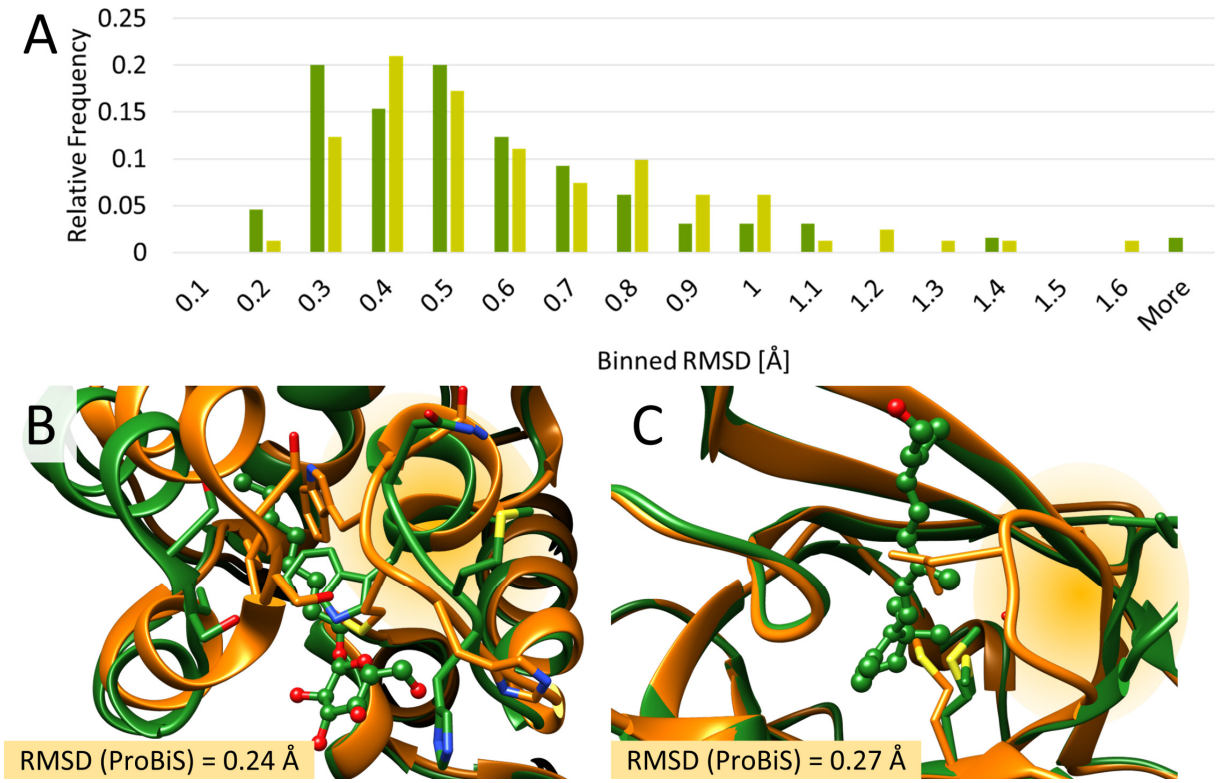


Figure 4.1.10. Comparing conformational changes upon ligand binding in the extended PDBbind core set and the dataset of cryptic site pairs. (A) The binned binding site RMSD values for the extended PDBbind core set (dark green bars) and cryptic site pairs (light green bars) as determined with the binding site comparison tool ProBiS are depicted. (B) The cryptic site pair for MAP kinase p38α is shown (green: PDB-ID 2npq in complex with a ligand, orange: PDB-ID 2zb1 with the unoccupied lipid binding site). The site conformations are very different although this difference is not reflected by the ProBiS-derived RMSD value. (C) Differences in the binding sites of β-lactoglobulin in the presence (green, PDB-ID 1gx8) and absence (orange, PDB-ID 1bsq) of ligands. The movement of a loop in the proximity of the binding site leads to the partial closure of the cavity in the apo state.

TM-SITE is the most robust (a significantly worse *RRO* for the apo site was observed in only 17% of all pairs) and well-performing method for this dataset, together with the ConCavity versions considering residue conservation and P2RANK. MetaPocket 2.0 has an overall very poor performance for this dataset. Except for SURFNET, the underlying tools PASS, ConCavity^{LIGSITE}, fpocket, LIGSITE^{CS}, and ghecom perform much better than the corresponding meta approach.

It seems likely that knowledge-based and combined methods outperform geometry- and energy-based approaches for this dataset. However, this cannot be observed. Although most geometry-based methods have hit rates below 20% for the unbound sites, the same holds true for the knowledge-based and combined methods. A plausible explanation is the overall poor performance of these methods irrespective of the binding site state (bound/unbound). A potential reason might be the uniqueness of the assembled structure pairs. Representative templates might be missing for knowledge-based methods. Additionally, COFACTOR, TM-SITE, S-SITE, and G-LoSA were applied excluding protein structures with high sequence similarity to the query (only template structures with a sequence identity of at most 30% were considered). Their inclusion might lead to improved performance for the four tools.

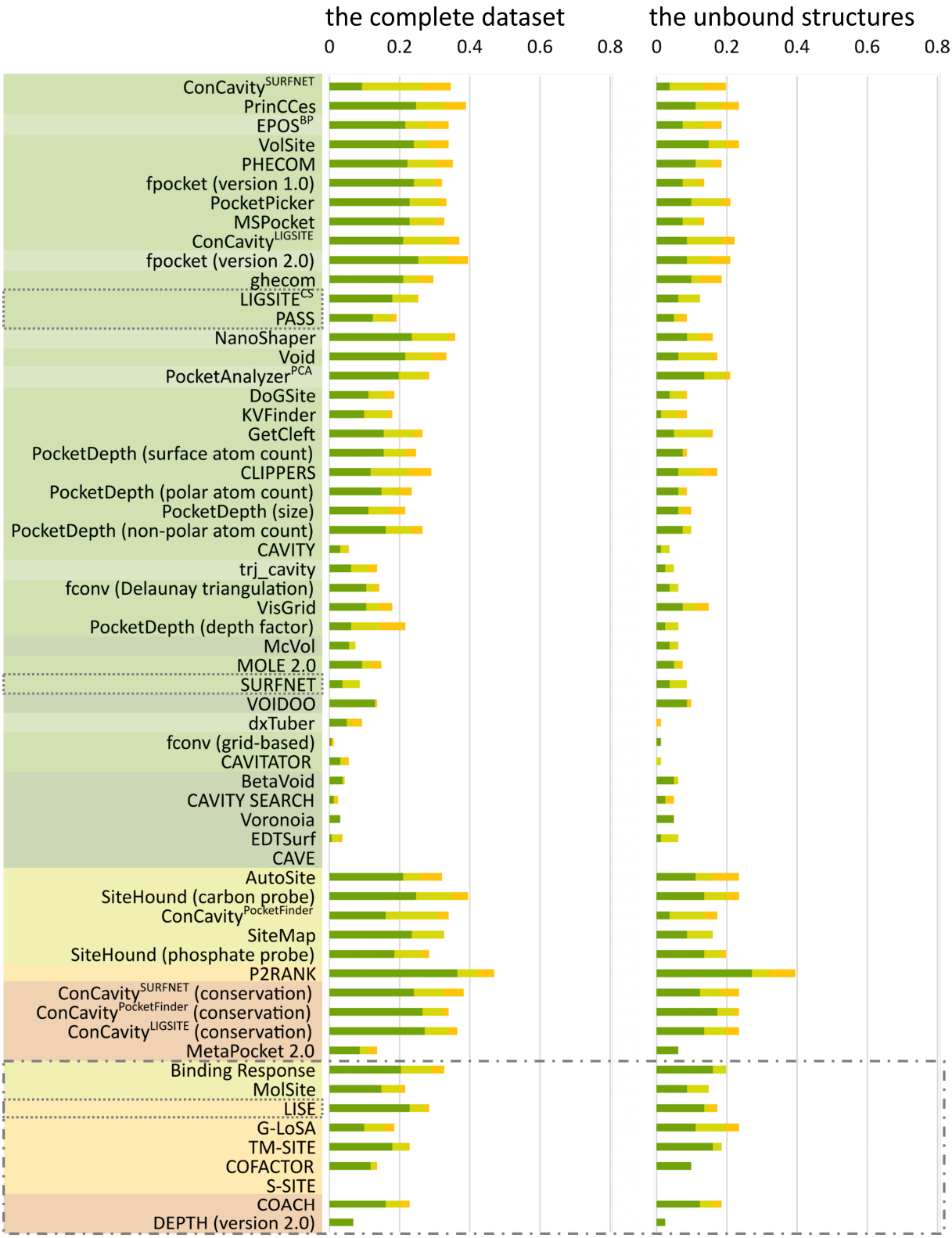


Figure 4.1.11. Residue-based evaluation of prediction tools for the cryptic binding site pairs. The top1 (dark green), top2 (light green), and top3 (yellow) pocket hit rates (*RRO* ≥ 0.5) of different binding site identification methodologies are given. The chart on the left gives the hit rates for the complete dataset, while the chart on the right represents the hit rates for the 81 apo structures. Methods highlighted in dark green, light green, yellow, and red predict binding sites based on geometry, energy, knowledge, and a combination thereof, respectively. The box (dot-dashed) highlights tools which are computationally very expensive. Dashed boxes highlight methods which annotate predicted sites by a single point. The light and dark grey boxes are used for tools that can be applied to MD trajectories and to detect interior protein binding sites, respectively.

The outstanding tool for this dataset is the knowledge-based method P2RANK, which correctly predicts nearly 40% of the unbound sites. The method is based on a machine learning technique for pocket scoring. The dataset of Chen and co-workers was used to train the RF classification model which is applied to predict the "ligandability" of protein pocket surface points.[223] The overlap of the structures in the benchmark datasets with those of the dataset of Chen *et al.*[43] was analysed to exclude that the evaluation was biased towards structures which were used for the RF training (Figure 4.1.12). We observe an overlap of 0.38%, 17.5%, and 28% with respect to structure, protein, and enzyme class, respectively, with the extended PDBbind core set. The cryptic sites and the dataset of Chen *et al.* overlap by 1.85%, 17.28%, and 24.56% with respect to structure, protein, and enzyme class. These overlaps do not explain the superior performance of P2RANK herein. Although some bias exists (at least with respect to proteins and enzyme classes), the external predictiveness of P2RANK independent of the applied dataset is evident.

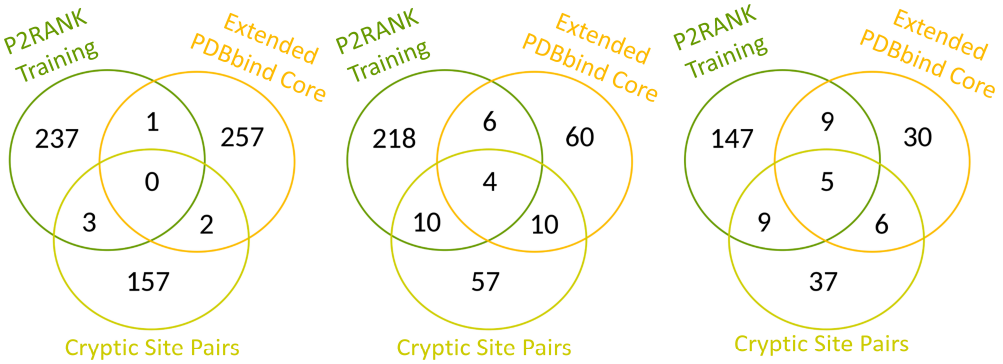


Figure 4.1.12. Similarities between the benchmark datasets and the dataset used for the development of P2RANK. The Venn diagrams show the overlap between the dataset of Chen *et al.* (dark green, used for the RF classifier training for P2RANK), the extended PDBbind core set (yellow), and the set of cryptic site pairs (light green) with respect to the structures, i.e., the PDB-IDs (left), the proteins, i.e., the UniProt accession numbers (centre), and the EC numbers (right).

The *RLO*-based hit rates allow for the even better discrimination between poorly performing and promising tools (Figure 4.1.13). P2RANK still outperforms the remaining tools (especially with respect to the top1 pocket hit rates), but other tools (ConCavity^{LIGSITE} (conservation), ConCavity^{SURFNET} (conservation), PocketPicker) perform equally good regarding the top3 pockets. The inclusion of conservation data for ConCavity pocket predictions significantly improved the top1 hit rate of ConCavity^{SURFNET}, but the hit rates of ConCavity^{LIGSITE} and ConCavity^{PocketFinder} were unaffected as already observed for the first dataset. P2RANK is nearly 10% (12%) better than the second-best tool if only the top3 (top1) ranked pockets for the apo sites are used for evaluation purposes. Only one geometry-based method, CAVITY, leads to comparable hit rates which are only slightly influenced by the presence or absence of a ligand. Although the energy-based methods perform good on the complete set, we observe a drop in the hit rates (of both probe- and docking-based approaches) by approximately 50% if only the apo sites are considered. Consequently, knowledge-based and combined approaches are most appropriate to predict challenging sites. Again, the performance of the combined approaches MetaPocket 2.0 and COACH is often worse than that of the methods whose results are combined. For COACH, this can be explained by the poor performing method S-SITE. For MetaPocket, the low hit rates of PASS and SURFNET might explain this trend.

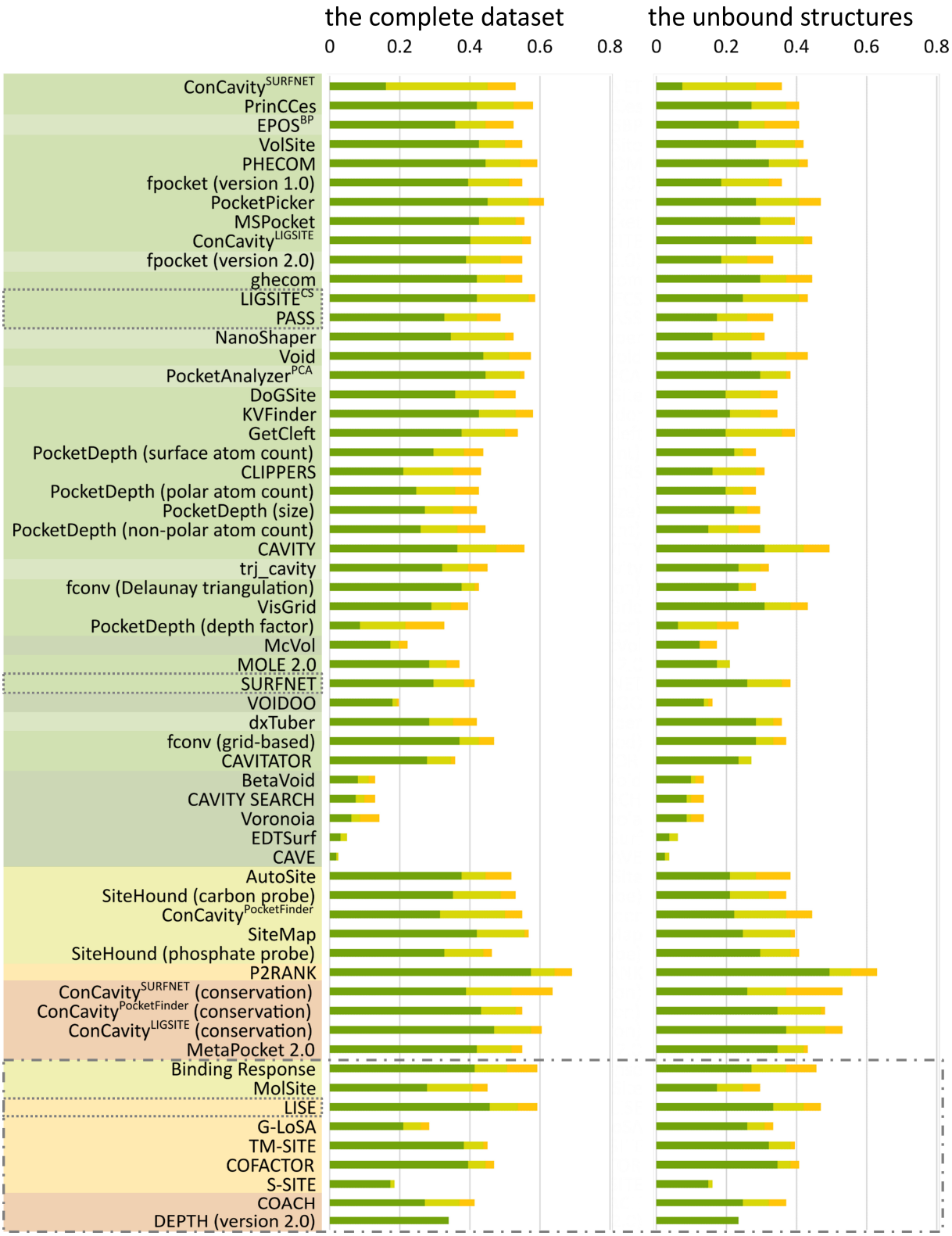


Figure 4.1.13. Ligand-based evaluation of prediction tools for the cryptic binding site pairs. The top1 (dark green), top2 (light green), and top3 (yellow) pocket hit rates (*RLO* ≥ 0.5) of different binding site identification methodologies are presented. The chart on the left shows the hit rates for the complete dataset, while the chart on the right represents the hit rates for the 65 apo structures. Methods highlighted in dark green, light green, yellow, and red predict binding sites based on geometry, energy, knowledge, and a combination thereof, respectively. The box (dot-dashed) highlights tools which are computationally very expensive. Dashed boxes highlight methods which annotate predicted sites by a single point. The light and dark grey boxes are used for tools that can be applied to MD trajectories and to detect interior protein binding sites, respectively.

Finally, the accuracy was assessed by analysing the hit rates for different *RLO* thresholds (Figure 4.1.14). P2RANK outperforms all tools irrespective of the applied hit threshold. Nevertheless, we can observe an adjustment of the hit rates with increasing *RLO* thresholds. While P2RANK shows clearly better performance at the thresholds 0.5, 0.6, and 0.7, only small differences (less than 2.5%) between P2RANK and the methods PrinCCes, ConCavity^{SURFNET} (conservation), VolSite, and LISE are observed at an *RLO* threshold of 0.8. Additionally, AutoSite, PocketPicker, and EPOS^{BP} show similar hit rates at a threshold of 0.9. So, if interested in very accurately predicted binding hot spots or the analysis of MD-derived ensembles, the user can expect equally reliable site predictions by AutoSite or EPOS^{BP}.

Interestingly, ConCavity^{SURFNET} (conservation) performed well for the structures of the PDBbind core set and even better than P2RANK with respect to the *RRO*. With respect to the cryptic sites, Its performance decreases. The same decline is observed for both versions of fpocket. A rescoring of all pockets predicted by different versions of ConCavity and fpocket was performed with PRANK. The results are presented in Table 4.1.5. The top3 and top1 hit rates of ConCavity improved only slightly (if at all) after rescoring with P2RANK. In contrast, hit rates of fpocket could be increased by up to 8%. The highest increase was observed for fpocket (version 1.0). However, none of the tools' predictions is as reliable as those of P2RANK. The rescoring of the ConCavity^{LIGSITE} predictions even led to decreased hit rates.

Table 4.1.5. Pocket rescoring with PRANK for the set of cryptic site pairs. The top3 (top1) hit rates (based on *RRO* and *RLO*) for the cryptic site pairs of pocket predictions before and after rescoring with PRANK are given. The hit rate for P2RANK is additionally given for comparison purposes. The difference between the hit rate after rescoring and the initial hit rate is given (Δ) and highlighted in bold characters if a hit rate increase by more than 5% was observed.

Method	Hit Rate for an *RRO* ≥ 0.5			Hit Rate for an *RLO* ≥ 0.5		
	Original	***After Rescoring***	***Δ***	***Original***	***After Rescoring***	***Δ***
fpocket (version 1.0)	0.321 (0.241)	0.395 (0.290)	**0.074** (0.049)	0.549 (0.395)	0.630 (0.494)	**0.080** (**0.099**)
fpocket (version 2.0)	0.395 (0.253)	0.420 (0.302)	0.025 (0.049)	0.549 (0.389)	0.611 (0.451)	**0.062** (**0.062**)
ConCavity^{LIGSITE}	0.370 (0.210)	0.352 (0.167)	-0.019 (-0.043)	0.574 (0.401)	0.556 (0.315)	-0.019 (-0.086)
ConCavity^{LIGSITE} (conservation)	0.364 (0.272)	0.364 (0.253)	0.000 (-0.019)	0.605 (0.469)	0.611 (0.438)	0.006 (-0.031)
ConCavity^{PocketFinder}	0.340 (0.160)	0.352 (0.216)	0.012 (**0.056**)	0.549 (0.315)	0.556 (0.352)	0.006 (0.037)
ConCavity^{PocketFinder} (conservation)	0.340 (0.265)	0.340 (0.265)	0.000 (0.000)	0.549 (0.432)	0.556 (0.438)	0.006 (0.006)
ConCavity^{SURFNET}	0.346 (0.093)	0.346 (0.136)	0.000 (0.043)	0.531 (0.160)	0.531 (0.241)	0.000 (**0.080**)
ConCavity^{SURFNET} (conservation)	0.383 (0.241)	0.383 (0.173)	0.000 (-0.068)	0.636 (0.389)	0.636 (0.333)	0.000 (-0.056)
P2RANK	0.469 (0.364)			0.691 (0.574)		

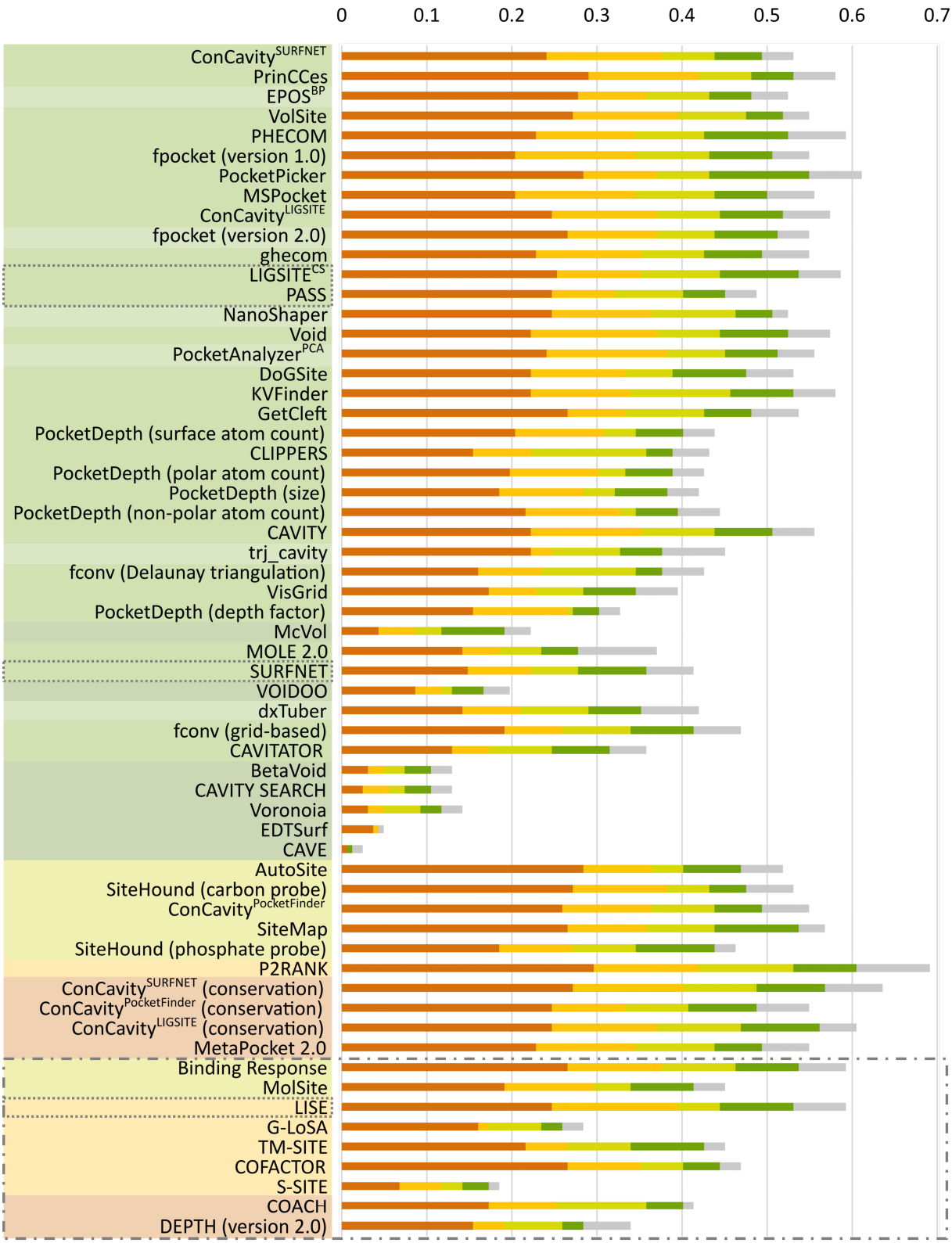


Figure 4.1.14. Threshold dependence of the ligand-based evaluation of prediction tools for the set of cryptic site pairs. The top3 pocket hit rates for an *RLO* of at least 0.5 (grey), 0.6 (dark green), 0.7 (light green), 0.8 (yellow), and 0.9 (red) of different binding site identification methodologies are shown. Methods highlighted in dark green, light green, yellow, and red predict binding sites based on geometry, energy, knowledge, and a combination thereof, respectively. The box (dot-dashed) highlights tools which are computationally very expensive. Dashed boxes highlight methods which annotate predicted sites by a single point. The light and dark grey boxes are used for tools that can be applied to MD trajectories and to detect interior protein binding sites, respectively.

Although P2RANK correctly predicts 69% of the cryptic binding sites (*RLO* ≥ 0.5), its performance is much worse than for the PDBbind core dataset (a hit rate of 95% for an *RLO* ≥ 0.5 was observed). The number of correctly predicted sites slightly increases to 75% after the exclusion of the unbound structures. Nevertheless, it is worthwhile analysing the challenging binding sites in more detail.

The number of tools per structure which did not correctly identify the binding site of interest was calculated. Altogether, 12 binding sites could not be identified by any of the tools. For more than a quarter of the dataset structures (27%), only 10 of the 60 tested tools succeeded in correctly predicting the site. Calculating the difference in the number of tools that identified the bound and unbound binding site per site pair, difficult structures due to binding site changes upon ligand binding can be identified. Of the 81 structure pairs, only 12 pairs were characterised by high differences in the performance of the methods. The binding site conformations are obviously not the only challenge of the dataset.

In comparison to the extended PDBbind core set, some ligands of the holo structures are characterised by comparatively low affinities (e.g., K_d = 160 µM for pantoate in the structure with the PDB-ID 2ofp, K_d = 550 µM for uric acid in the structure with the PDB-ID 1l5s). As many of the methods were trained and tested for high-affinity protein-ligand complexes, such cases might become challenging to predict. Table A1.5 provides all available affinities and inhibitory constants for the ligands of this dataset.

The DoGSite descriptors of all binding sites of the ligand-bound structures were calculated (see Tables A1.4 and A1.7). Analysing the results, two further reasons for the failure of the tools to detect all sites of interest became evident. First, the cryptic site pair set includes very small binding sites in complex with low molecular weight ligands (e.g., pantoate, phenylacetaldehyde, or uric acid). In the dataset, 11% of the protein structures are bound to ligands with at most 20 heteroatoms. In contrast, none of the binding site ligands in the extended PDBbind core set has 20 or even fewer heteroatoms. This leads to a low ranking of the corresponding sites of the cryptic pairs based on surface or volume as realised by many of the geometry-based tools. The additive nature of some scores for the knowledge-based and combined methods will also lead to a low ranking of the respective sites. Second, the average lid area was larger for the cryptic binding sites (132 Å² vs. 103 Å²) which hints at flat solvent-exposed sites in the protein structures. Furthermore, sugar binding sites, which were also rarely ranked in the top3 pockets for the PDBbind core set, were not correctly predicted and lead to the decrease in the overall hit rates.

4.1.8 Preliminary Conclusions

Finally, the results for both datasets and many quality indicators have to be summarised to enable the choice of a reliable tool. This decision depends on the resources, the number of protein structures to be analysed, the way the extracted pockets are post-processed, and many other criteria. Therefore, it is not possible to assign “the best performing tool”. Instead, important standards are specified which can assist in the choice of the most appropriate tool. For this purpose, the performance of the methods was grouped into three categories (good, intermediate, and poor performance) per evaluation criterion (Figure 4.1.15). The thresholds were set separately for all criteria based on the individual distributions of the quality metrics for all methods.

Some examples shall illustrate the usage of Figure 4.1.15 as guidance for the choice of an appropriate tool. If a druggability assessment is required for predicted sites, CAVITY seems to be the best choice. Although fpocket (version 2.0) and DoGSite have higher top3 hit rates for the apo sites of the PDBbind core set, they show lower robustness for the cryptic site pairs and their performance is poor for the challenging cases of the datasets.

The analysis of the pocket dynamics by means of MD trajectories is possible with the most reliable tools PocketAnalyzer^{PCA} or fpocket (version 2.0). If small or very flexible sites are the focus of the study, fpocket (version 2.0) should be the tool of choice. PocketAnalyzer^{PCA} is biased towards pockets with a low lid-to-hull ratio, i.e., highly enclosed or buried sites and failed to correctly predict the challenging sites in the datasets.

Talking about difficult cases, it should also be highlighted that especially slow tools and knowledge-based methods identify sites which were not predicted by most geometry-based approaches. They are worth using for scarcely characterised proteins and to identify novel biologically relevant pockets in protein structures, but restricted to small numbers of input structures.

For the processing of large datasets, P2RANK is the best choice. It led to good results for all but one of the quality criteria presented. It showed intermediate, but not poor performance for the highly flexible and small pockets of the cryptic site pairs set. A combination of P2RANK with other knowledge-based methods might lead to further improvements.

Finally, Figure 4.1.15 hints at the last especially interesting aspect when dealing with a large number of prediction methodologies. The advantages and disadvantages of some methods are complementary. DoGSite shows a good performance considering the top3 ranked pockets, but a bad performance for the challenging sites. G-LoSA and TM-SITE proved as especially useful for sites which other methods miss and might perfectly supplement this method. Probably, combined approaches might lead to more promising results if a more appropriate combination of approaches was used. An inspection of the overall performance of ConCavity, MetaPocket 2.0, COACH, and DEPTH (version 2.0) indicates the necessity for and chance of optimisations of the already available meta approaches.

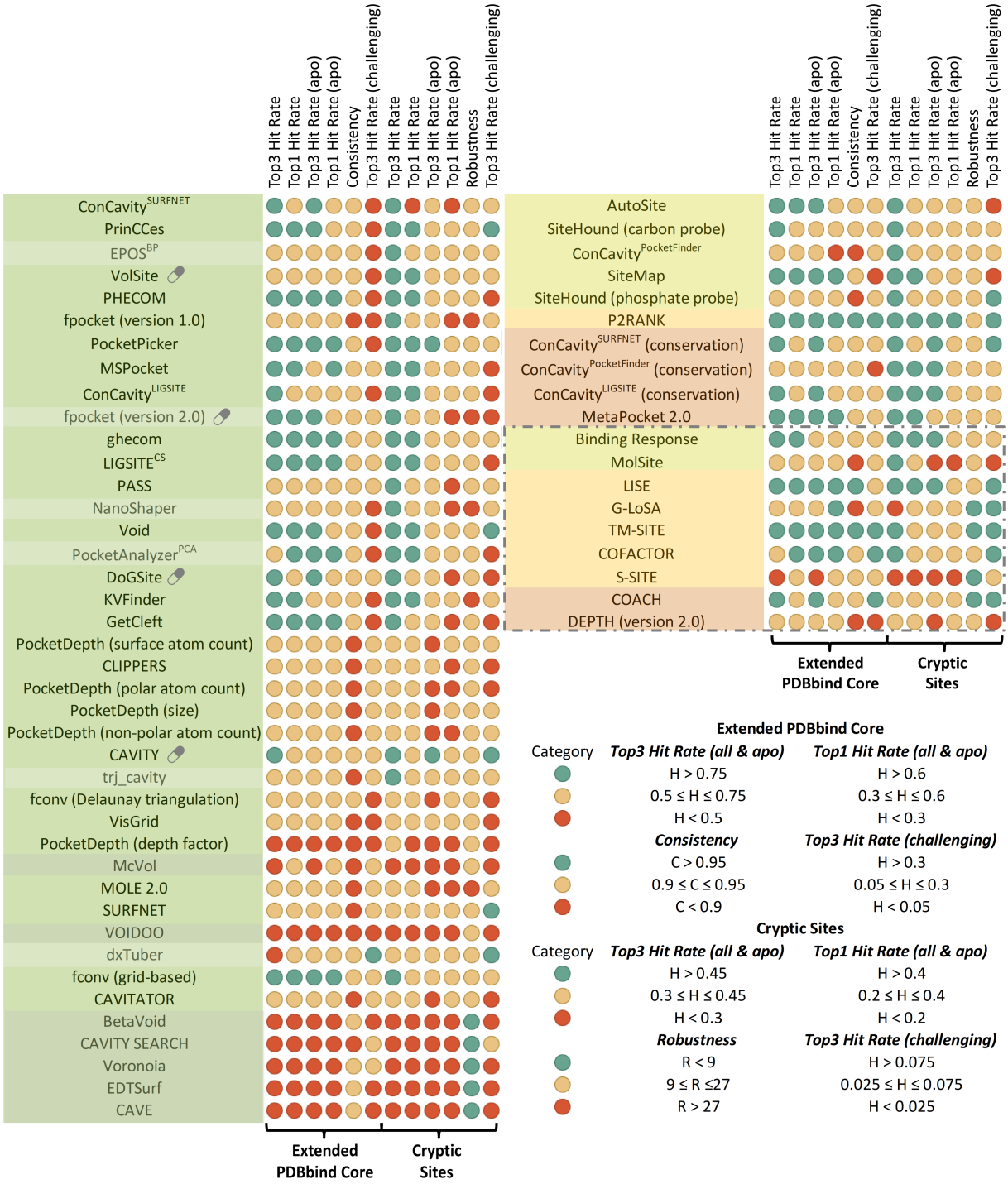


Figure 4.1.15. The binding site prediction tools' performance based on different evaluation criteria. For both datasets, the top3 and top1 hit rates for the complete dataset and the apo structures alone were considered. The top3 *RLO*-based hit rate for the challenging sites (which were not correctly identified by at least 51 of the 60 evaluated methods based on an *RLO* ≥ 0.5) was calculated. The consistency within the groups of identical proteins of the extended PDBbind core set was evaluated. The robustness with respect to conformational variations for the cryptic site pairs was evaluated by counting all pairs with an *RLO* difference between the holo and apo site above 0.5 per method (the lower the number, the lower the tools' sensitivity). The light and dark grey boxes are used for tools that can be applied to MD trajectories and to detect interior protein binding sites, respectively. The pill symbol highlights the tools which also assess the druggability of the binding sites under investigation.

4.1.9 Binding Site Prediction for Homology Models and Allosteric Sites

After examining the difficult cases within the previous datasets, it is also interesting to investigate further challenging sites referred to in the literature. First, it was analysed whether the hit rates decrease for structures that were created by comparative modelling. A drop in prediction accuracy was previously observed, although some methods (often knowledge-based ones) lack this dependence on the type of structure (experimental vs. homology model).[36,220] A set of ten *M. tuberculosis* protein structures, which were modelled based on templates with sequence similarities not exceeding 40%, was used. The template structures were also included. The templates are bound to 11 ligands, leading to a dataset of 22 binding sites. This dataset enabled the comparison between the hit rates for experimentally derived and theoretical models for all methods. The results are presented in Table 4.1.6.

Table 4.1.6. Binding site prediction performance for comparative models. The top3 hit rates (*RLO* or *RRO* ≥ 0.5) of all methods for the comparative models and their templates are compared. The difference between the two hit rates (Δ) is also given. A colour gradient from green (small Δ) to yellow to red (high Δ) was applied. The numbers in parentheses for ConCavity correspond to the hit rates if HSSP files are generated for the sequences of the homology models.

Method	*RRO*			*RLO*		
	Template Structures	***Homology Models***	***Δ***	***Template Structures***	***Homology Models***	***Δ***
ConCavity^{SURFNET}	0.36	0.27	0.09	0.55	0.45	0.09
PrinCCes	0.27	0.27	0.00	0.55	0.55	0.00
EPOS^{BP}	0.45	0.18	0.27	0.55	0.45	0.09
VolSite	0.27	0.45	-0.18	0.64	0.55	0.09
PHECOM	0.36	0.27	0.09	0.55	0.64	-0.09
fpocket (version 1.0)	0.36	0.27	0.09	0.64	0.45	0.18
PocketPicker	0.36	0.27	0.09	0.55	0.45	0.09
MSPocket	0.27	0.36	-0.09	0.27	0.27	0.00
ConCavity^{LIGSITE}	0.45	0.27	0.18	0.73	0.64	0.09
fpocket (version 2.0)	0.27	0.18	0.09	0.55	0.45	0.09
ghecom	0.45	0.45	0.00	0.64	0.55	0.09
LIGSITE^{CS}	0.18	0.18	0.00	0.45	0.55	-0.09
PASS	0.36	0.27	0.09	0.64	0.55	0.09
NanoShaper	0.27	0.18	0.09	0.55	0.45	0.09
Void	0.36	0.27	0.09	0.45	0.36	0.09
PocketAnalyzer^{PCA}	0.27	0.27	0.00	0.64	0.45	0.18
DoGSite	0.27	0.18	0.09	0.45	0.45	0.00
KVFinder	0.18	0.18	0.00	0.45	0.36	0.09
GetCleft	0.27	0.27	0.00	0.55	0.45	0.09
PocketDepth (surface atoms)	0.18	0.09	0.09	0.36	0.27	0.09
CLIPPERS	0.27	0.09	0.18	0.45	0.27	0.18
PocketDepth (polar atoms)	0.18	0.09	0.09	0.27	0.18	0.09
PocketDepth (size)	0.09	0.00	0.09	0.36	0.18	0.18
PocketDepth (non-polar atoms)	0.18	0.18	0.00	0.27	0.27	0.00
CAVITY	0.09	0.18	-0.09	0.45	0.36	0.09

Table 4.1.6 *(continued)*. Binding site prediction performance for comparative models. The top3 hit rates (*RLO* or *RRO* ≥ 0.5) of all methods for comparative models and their templates are compared. The difference between both hit rates (Δ) is also given. A colour gradient from green (small Δ) to yellow to red (high Δ) was applied. The numbers in parentheses for ConCavity correspond to the hit rates if HSSP files are generated for the sequences of the homology models.

Method	*RRO*			*RLO*		
	Template Structures	*Homology Models*	*Δ*	*Template Structures*	*Homology Models*	*Δ*
trj_cavity	0.09	0.27	-0.18	0.45	0.18	0.27
fconv (Delaunay triangulation)	0.18	0.18	0.00	0.36	0.55	-0.18
VisGrid	0.27	0.27	0.00	0.36	0.36	0.00
PocketDepth (depth factor)	0.00	0.09	-0.09	0.18	0.18	0.00
McVol	0.18	0.18	0.00	0.36	0.36	0.00
MOLE 2.0	0.18	0.27	-0.09	0.27	0.27	0.00
SURFNET	0.09	0.09	0.00	0.36	0.27	0.09
VOIDOO	0.18	0.09	0.09	0.27	0.18	0.09
dxTuber	0.00	0.00	0.00	0.18	0.27	-0.09
fconv (grid-based)	0.09	0.09	0.00	0.36	0.45	-0.09
CAVITATOR	0.00	0.00	0.00	0.36	0.27	0.09
BetaVoid	0.00	0.09	-0.09	0.18	0.18	0.00
CAVITY SEARCH	0.09	0.09	0.00	0.27	0.27	0.00
Voronoia	0.00	0.09	-0.09	0.09	0.18	-0.09
EDTSurf	0.00	0.09	-0.09	0.09	0.18	-0.09
CAVE	0.00	0.00	0.00	0.00	0.00	0.00
AutoSite	0.36	0.27	0.09	0.64	0.45	0.18
SiteHound (carbon probe)	0.36	0.18	0.18	0.45	0.27	0.18
ConCavity^{PocketFinder}	0.45	0.27	0.18	0.64	0.45	0.18
SiteMap	0.27	0.27	0.00	0.64	0.55	0.09
SiteHound (phosphate probe)	0.27	0.45	-0.18	0.64	0.73	-0.09
P2RANK	0.45	0.27	0.18	0.73	0.64	0.09
ConCavity^{SURFNET} (conservation)	0.55	0.27 (0.18)	0.27 (0.36)	0.73	0.45 (0.36)	0.27 (0.36)
ConCavity^{PocketFinder} (conservation)	0.45	0.09 (0.00)	0.36 (0.45)	0.64	0.36 (0.18)	0.27 (0.45)
ConCavity^{LIGSITE} (conservation)	0.27	0.18 (0.27)	0.09 (0.00)	0.73	0.18 (0.45)	0.55 (0.27)
MetaPocket 2.0	0.18	0.27	-0.09	0.64	0.64	0.00
Binding Response	0.64	0.18	0.45	0.64	0.36	0.27
MolSite	0.36	0.09	0.27	0.64	0.55	0.09
LISE	0.27	0.27	0.00	0.55	0.55	0.00
G-LoSA	0.18	0.27	-0.09	0.45	0.27	0.18
TM-SITE	0.18	0.27	-0.09	0.45	0.55	-0.09
COFACTOR	0.27	0.36	-0.09	0.73	0.64	0.09
S-SITE	0.00	0.00	0.00	0.45	0.18	0.27
COACH	0.27	0.36	-0.09	0.64	0.55	0.09
DEPTH (version 2.0)	0.00	0.00	0.00	0.36	0.36	0.00

At first glance, the hit rates of none of the tools drastically decrease for the homology models in this dataset. Nevertheless, especially energy-based methods fail to identify the correct binding sites of the theoretical models, although they correctly predict the template binding sites. SiteMap and SiteHound (using a phosphate probe) are the only exceptions. A significant correlation between the RMSD for the alignment of the template and model structures and the *RLO*-based top3 hit rate was also observed for PASS (see Table A1.10 for the Pearson correlation coefficients for all tools).

The drop in performance for the conservation-based ConCavity algorithms can be partially attributed to the missing residue conservation information (HSSP files) for the models. However, the corresponding methods without conservation data also identify fewer binding sites for the homology models. The xssp-server[349] was used to generate the HSSP files for the theoretical models. The *RRO*-based top3 hit rates for ConCavity^{LIGSITE} improved (0.18 to 0.27). The same holds true for the *RLO*-based hit rates (0.18 to 0.45). Intriguingly, the hit rates for ConCavity using the PocketFinder or SURFNET methodology decreased. Most knowledge-based approaches, but especially LISE, do not show huge prediction differences for the sites of experimentally solved structures and the corresponding *M. tuberculosis* models. SiteHound (using a phosphate probe), COFACTOR, ConCavity^{LIGSITE}, P2RANK, MetaPocket 2.0, and PHECOM led to the highest top3 hit rates (*RLO* ≥ 0.5) for the homology models.

Allosteric pockets are often also difficult to predict. Given the promising performance of some methods on the previous datasets, a set of allosteric sites in proteins of a diverse dataset was used to evaluate the prediction performance (Figure 4.1.16). Nearly all methods show a huge difference in the top1 and top3 hit rates. As most tools were developed to predict active sites, the allosteric pockets show lower scores in most cases, i.e., they appear lower in the ranking list than active sites and cofactor-binding pockets. Although the dataset comprises only structures with bound ligands, the percentage of correct predictions significantly drops when compared to that for the ligand-bound structures of the previous datasets. The main reason is the different nature of the allosteric sites compared to the orthosteric ones. Most allosteric ligands bind at interfaces of multiple protein chains. Using multimeric protein structures for binding site prediction leads to a high number of predicted pockets. In consequence, the reliable scoring of the detected pockets becomes the major determinant of the prediction performance. Although P2RANK-based pocket ranking was successful for the extended PDBbind core set and the cryptic sites, a ranking according to the pocket volume as realised with LIGSITE^{CS} outperforms the former method. This hints at a lack of allosteric sites in the dataset used for the training of the RF-based approach P2RANK. Intriguingly, PocketAnalyzer^{PCA} showed the best performance for the top1, top2, and top3 ranked pockets, followed by LIGSITE^{CS} and PrinCCes. SiteMap and P2RANK led to the highest hit rates within the energy- and knowledge-based methods, respectively. Nevertheless, it is obvious that a prediction tool specially developed for allosteric sites might be a better choice for this dataset. This assumption was tested by applying the tool ALLO[230] to rescore the sites that were predicted by DoGSite. This led to an increase in the *RRO*-based and *RLO*-based top3 hit rates from 0.18 to 0.25 and 0.41 to 0.47, respectively.

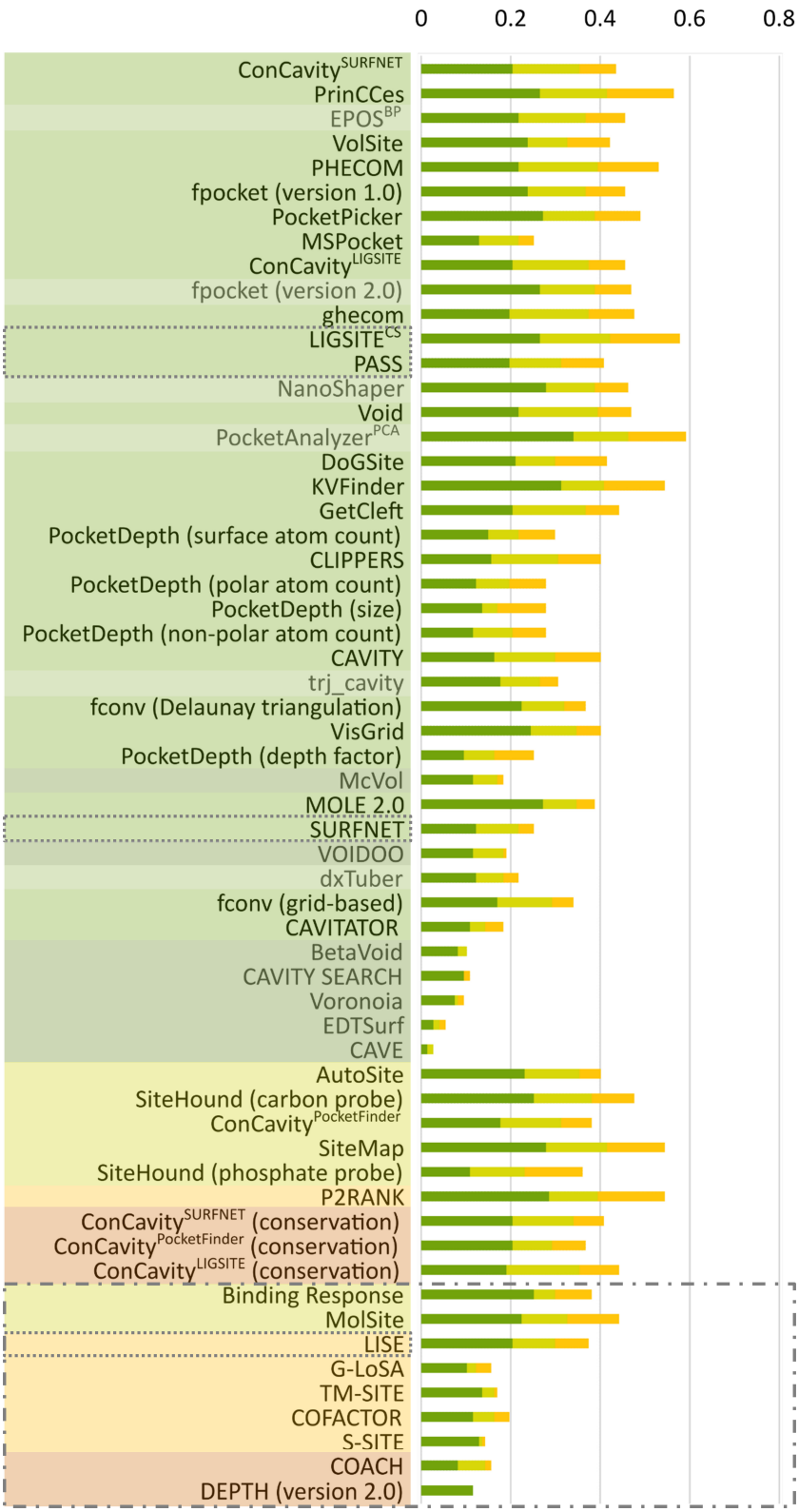


Figure 4.1.16. Ligand-based evaluation of prediction tools for the dataset of allosteric sites. The top1 (dark green), top2 (light green), and top3 (yellow) pocket hit rates (*RLO* ≥ 0.5) of the different binding site identification methodologies are given. Methods highlighted in dark green, light green, yellow, and red predict binding sites based on the geometry, energy, similarity, and the combination thereof, respectively. The box (dot-dashed) highlights tools which are computationally very expensive. Dashed boxes highlight methods which annotate predicted sites by a single point. The light and dark grey boxes are used for tools that can be applied to MD trajectories and to detect interior protein binding sites, respectively.

As shown previously[350], these investigations highlight that very specialised scoring schemes are necessary to favour the allosteric sites in the ranked lists of predicted pockets. Consequently, the top1 and top3 hit rates for the allosteric sites are insufficient to analyse the methods' ability to detect allosteric sites. The *RLO*-based top4 to top10 hit rates and the hit rates if all predicted cavities are considered were calculated to evaluate whether the identification or the scoring of the detected binding sites leads to the poor performance of all tools. As presented in Table 4.1.7, the allosteric sites were correctly identified by most tools, but they were ranked considerably lower than other sites. EPOS^{BP} correctly identifies 90% of all binding sites. PrinCCes and MolSite predict 86% of all pockets. MolSite assigns Cα atoms across the complete protein surface as pockets for docking resulting in the high number of detected sites. Considering only the ten highest ranked pockets for all tools, PrinCCes and SiteHound (using a carbon probe) show the highest hit rates. Furthermore, EPOS^{BP}, PHECOM, fpocket (version 1.0), LIGSITE^{CS}, P2RANK, and Binding Response yield hit rates above 70%. The increasing hit rate with an increasing number of considered pockets, which is observed for most tools, indicates problems concerning the scoring of detected allosteric sites. Strikingly, most knowledge-based methods (the only exceptions are LISE and P2RANK) show very low hit rates for the allosteric sites even if all pockets are considered. Obviously, they fail to detect similar sites in the generated databases. This highlights the lack of allosteric sites in the datasets used to derive the predictions based on site similarities. Even the knowledge-based method P2RANK, which showed superior performance for all previously analysed datasets, is outperformed by several purely geometry-based approaches.

Table 4.1.7. The detection and scoring of allosteric binding sites. The *RLO*-based hit rates for the top*x* ranked pockets and all detected pockets are given per binding site prediction method. A colour gradient ranging from red (lowest hit rate) to yellow to green (highest hit rate) was applied.

Method	Top4	Top5	Top6	Top7	Top8	Top9	Top10	All
ConCavity^{SURFNET}	0.50	0.54	0.56	0.59	0.59	0.61	0.61	0.63
PrinCCes	0.61	0.69	0.71	0.76	0.76	0.76	0.77	0.86
EPOS^{BP}	0.54	0.58	0.63	0.68	0.69	0.72	0.73	0.90
VolSite	0.33	0.35	0.37	0.39	0.39	0.40	0.41	0.41
PHECOM	0.56	0.60	0.64	0.65	0.67	0.71	0.73	0.76
fpocket (version 1.0)	0.54	0.60	0.62	0.65	0.67	0.71	0.73	0.83
PocketPicker	0.57	0.63	0.65	0.67	0.67	0.68	0.69	0.79
MSPocket	0.29	0.33	0.33	0.34	0.35	0.35	0.36	0.41
ConCavity^{LIGSITE}	0.51	0.55	0.60	0.61	0.61	0.61	0.61	0.61
fpocket (version 2.0)	0.54	0.59	0.61	0.65	0.66	0.67	0.69	0.73
ghecom	0.51	0.53	0.59	0.62	0.64	0.66	0.67	0.71
LIGSITE^{CS}	0.63	0.67	0.70	0.71	0.73	0.74	0.74	0.83
PASS	0.47	0.51	0.56	0.60	0.63	0.64	0.65	0.74
NanoShaper	0.52	0.56	0.59	0.62	0.64	0.65	0.65	0.70
Void	0.52	0.56	0.59	0.62	0.63	0.64	0.66	0.66
PocketAnalyzer^{PCA}	0.61	0.64	0.65	0.67	0.67	0.67	0.67	0.68
DoGSite	0.49	0.53	0.56	0.57	0.60	0.61	0.63	0.70
KVFinder	0.59	0.61	0.63	0.65	0.66	0.66	0.69	0.71
GetCleft	0.50	0.52	0.54	0.58	0.59	0.60	0.61	0.74
PocketDepth (surface atoms)	0.37	0.41	0.45	0.51	0.56	0.57	0.59	0.75
CLIPPERS	0.44	0.47	0.48	0.49	0.49	0.49	0.49	0.50

Table 4.1.7 *(continued)*. The detection and scoring of allosteric binding sites. The *RLO*-based hit rates for the top*x* ranked pockets and all detected pockets are given per binding site prediction method. A colour gradient ranging from red (lowest hit rate) to yellow to green (highest hit rate) was applied.

Method	Top4	Top5	Top6	Top7	Top8	Top9	Top10	All
PocketDepth (polar atoms)	0.34	0.38	0.43	0.48	0.54	0.54	0.59	0.75
PocketDepth (size)	0.33	0.39	0.42	0.46	0.48	0.54	0.57	0.75
PocketDepth (non-polar atoms)	0.37	0.41	0.47	0.52	0.54	0.57	0.60	0.75
CAVITY	0.47	0.52	0.58	0.61	0.62	0.63	0.63	0.68
trj_cavity	0.33	0.36	0.37	0.39	0.40	0.40	0.41	0.42
fconv (Delaunay triangulation)	0.37	0.38	0.38	0.38	0.38	0.38	0.39	0.39
VisGrid	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
PocketDepth (depth factor)	0.31	0.39	0.45	0.50	0.54	0.56	0.57	0.75
McVol	0.25	0.26	0.27	0.30	0.32	0.32	0.33	0.40
MOLE 2.0	0.43	0.44	0.45	0.46	0.46	0.46	0.46	0.47
SURFNET	0.28	0.29	0.29	0.32	0.33	0.35	0.36	0.42
VOIDOO	0.20	0.22	0.22	0.22	0.22	0.22	0.22	0.22
dxTuber	0.25	0.26	0.27	0.29	0.29	0.31	0.31	0.45
fconv (grid-based)	0.37	0.39	0.39	0.40	0.41	0.41	0.41	0.41
CAVITATOR	0.20	0.22	0.25	0.27	0.27	0.29	0.29	0.29
BetaVoid	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
CAVITY SEARCH	0.13	0.14	0.14	0.15	0.15	0.16	0.16	0.16
Voronoia	0.10	0.10	0.12	0.12	0.13	0.13	0.14	0.27
EDTSurf	0.05	0.07	0.07	0.09	0.09	0.09	0.10	0.10
CAVE	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
AutoSite	0.49	0.51	0.57	0.57	0.59	0.60	0.61	0.69
SiteHound (carbon probe)	0.52	0.59	0.63	0.67	0.71	0.73	0.76	0.76
ConCavity^{PocketFinder}	0.46	0.50	0.53	0.54	0.56	0.56	0.56	0.57
SiteMap	0.61	0.63	0.65	0.65	0.66	0.66	0.68	0.68
SiteHound (phosphate probe)	0.40	0.45	0.49	0.51	0.52	0.55	0.56	0.56
P2RANK	0.57	0.63	0.67	0.69	0.71	0.73	0.74	0.76
ConCavity^{SURFNET} (conservation)	0.49	0.57	0.59	0.59	0.60	0.60	0.61	0.61
ConCavity^{PocketFinder} (conservation)	0.40	0.44	0.46	0.46	0.46	0.46	0.46	0.46
ConCavity^{LIGSITE} (conservation)	0.47	0.51	0.51	0.51	0.51	0.51	0.52	0.52
Binding Response	0.46	0.52	0.59	0.61	0.65	0.67	0.73	0.73
MolSite	0.49	0.53	0.59	0.60	0.62	0.63	0.65	0.86
LISE	0.43	0.48	0.51	0.54	0.56	0.59	0.61	0.61
G-LoSA	0.18	0.20	0.22	0.22	0.23	0.24	0.24	0.33
TM-SITE	0.18	0.20	0.22	0.23	0.23	0.24	0.24	0.24
COFACTOR	0.22	0.23	0.25	0.25	0.26	0.26	0.29	0.31
S-SITE	0.14	0.14	0.14	0.15	0.15	0.15	0.15	0.15
COACH	0.19	0.20	0.22	0.22	0.23	0.23	0.23	0.28
DEPTH (version 2.0)	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12

Taken together, these results indicate two essential requirements for binding site detection. First, the performance of some methods depends on the type of the input structure (apo proteins, homology models). Second, the ranking of the detected pockets significantly influences the prediction hit rates. This necessitates a robust and generally applicable scoring function to enrich true positive binding sites. Especially knowledge-based methods might benefit from the inclusion of different site types (orthosteric, allosteric, sites at interfaces) in the datasets used for the detection of site similarities or the training of the method.

4.1.10 Establishment of a Meta Approach to Improve Prediction Performance

For the extended PDBbind core set and the cryptic site pairs, we saw that the performance of the MetaPocket approach is worse than expected. Many tools outperform this method which usually benefits from the complementarity of multiple binding site detection procedures. Therefore, novel combinations of reasonably fast and reliable methods, which might further improve the prediction performance, were tested.

The theoretical outcome of all possible combinations of all methods (combining up to eight methods) was assessed considering the method with the best top1 or top3 hit rate per pocket for each combination. The results are depicted in Figure 4.1.17 and guided the decision for methods to combine. The inclusion of computationally expensive tools did not lead to significantly higher hit rates with different combinations for the extended PDBbind core set. In all further analyses, these methods were not used to minimise the necessary time for the application of multiple methods. The hit rate differences between the optimum combinations based on the top1 and top3 ranked pockets show that a meta approach will most probably lead to the best results if the top3 pockets per method are considered. The hit rates reach a plateau of the theoretically achievable hit rate for combinations of more than five methods. In consequence, the optimum combinations of five methods that correctly predict the highest number of pockets were chosen based on the *RRO* and the *RLO* measure.

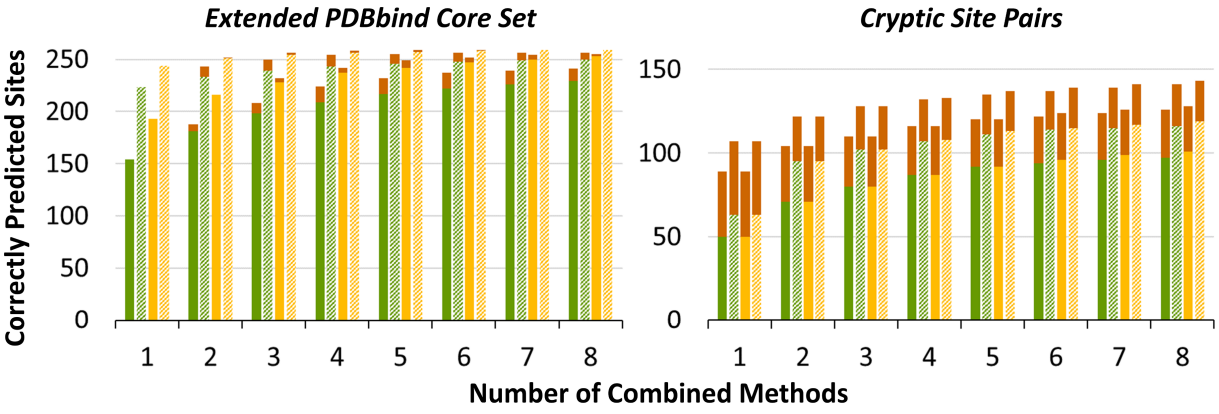


Figure 4.1.17. Theoretical numbers of correct predictions for the optimum method combinations. The theoretical numbers of correct predictions for combinations of up to eight methods are shown when considering the top1 (solid fill) and top3 (striped pattern) highest ranked pockets. The green and yellow bars show the number of correctly predicted sites if all sites with an *RRO* ≥ 0.5 and an *RLO* ≥ 0.5 are considered as correctly predicted, respectively. The results for the extended PDBbind core set (on the left) and the cryptic site pairs (on the right) are given. Red bars indicate the theoretically achievable hit rates if the computationally expensive tools are also included in the analysis.

Using the *RRO*-based hit rates, multiple combinations of the fast methods led to the correct predictions of 242 out of all 260 detectable sites. The theoretically achievable hit rate of all eight combinations for the dataset of cryptic sites was additionally calculated to identify the combination that leads to the optimum hit rates for both datasets. The combination of ConCavity^{LIGSITE} (conservation), ConCavity^{SURFNET} (conservation), P2RANK, SiteHound (using a carbon probe), and VolSite led to the maximum of 103 theoretically correctly predicted sites of all 164 cryptic sites.

The analysis based on the *RLO*-based hit rates led to 16 combinations which correctly identified 257 of all 260 sites in the extended PDBbind core set. Of these combinations, the inclusion of the top3 ranked pockets reported by dxTuber, MOLE 2.0, P2RANK, PocketDepth (ranked by the size of the sites), and Void led to the highest overall hit rate for the cryptic site pairs (129 of 164 sites were correctly predicted).

Both the *RRO*- and *RLO*-based optimum method combinations were subsequently tested for all analysed datasets. Using the top3 ranked pockets for a combination of 5 methods leads to 15 pockets. Duplicate pockets were iteratively filtered starting with the pockets identified by the most accurate method (based on the *RRO*-based top3 hit rate). The filtered sites were subsequently rescored using two different approaches.

In the first approach, the number of times a pocket was identified by the different methods was counted. Then, the pockets were ranked according to the number of methods which ranked them within the top3 pockets. If this number was equal for two pockets, the pockets were ranked according to the performance of the methods which predicted them. This leads to a ranking list based on the "popularity" of the predicted sites (FAVOUR). In the second approach, the filtered pockets were rescored with PRANK and the resulting rankings were considered to identify the three highest ranked pockets per protein (PRANK).

The results for both analysed combinations and the rescoring procedures are presented in Table 4.1.8. For the prediction accuracy including proper binding site dimensions and distinct residue assignments (quality measures: *RRO* ≥ 0.5 and *RLO* = 1), the *RRO*-based combination of $\text{ConCavity}^{\text{LIGSITE}}$ (conservation), $\text{ConCavity}^{\text{SURFNET}}$ (conservation), P2RANK, SiteHound (carbon probe), and VolSite seems to be the best choice. The rescoring with PRANK is the most reliable scoring option in this case. In contrast, the quality measure *RLO* ≥ 0.5 indicates the *RLO*-based combination (dxTuber, MOLE 2.0, P2RANK, PocketDepth (size), Void) as the most predictive one. In most cases, this holds only true if the FAVOUR method is used for the pocket ranking. However, the overall improvements are most pronounced for the extended PDBbind core dataset and a comparison to the predominantly best performing method P2RANK shows that, similar to COACH and MetaPocket 2.0 in the benchmark, the meta approach does not necessarily lead to major improvements. This is especially interesting as the top3 pockets of P2RANK are considered in both combinations. What is more, we find lower top1 pocket hit rates for both approaches although the top3 hit rates are similar to those of P2RANK.

These findings point towards the still challenging issue concerning the scoring of the predicted sites. Although it is known that the 15 pockets predicted by the combinations of the methods contain the correct one in 257 of 260 cases for the PDBbind core set, the maximum hit rates of both combinations (0.90 and 0.92) do not reach the theoretically predicted hit rate of 0.99. A more elaborate scoring procedure might be the key to this problem. The superior performance of P2RANK for three of the datasets and the decrease in the performance for the less studied allosteric sites shows that an extension of this method using an even larger or more diverse (with respect to the binding site nature) dataset for training purposes might provide a useful means to meet this final challenge of binding site prediction.

Nevertheless, the results for the more restrictive quality measures show that an increase in the accuracy of the predicted sites is expectable. This is the basic benefit of the meta approaches which were established and analysed herein. Further investigations have to pave the way to such extraordinarily high top3 hit rates as 0.99.

Table 4.1.8. Results of two meta approaches for the prediction of binding sites. The top1, top2, and top3 hit rates of two meta approaches (*RRO*-based and *RLO*-based combination) are shown for the rescoring methodologies applied in this study (FAVOUR/PRANK). For comparison purposes, the P2RANK hit rates are also given in italics. The meta approaches were applied to all benchmark datasets applied herein. The colour gradient ranges from red (lowest hit rate) to yellow to green (highest hit rate). The highest top3 hit rates per performance measure are given in bold characters.

Dataset	Performance Measure	Rescoring	***RRO*-Based Combination**			***RLO*-Based Combination**		
			Top1	***Top2***	***Top3***	***Top1***	***Top2***	***Top3***
Extended PDBbind Core	*RRO* ≥ 0.5	*P2RANK*	*0.62*	*0.65*	*0.67*	*0.62*	*0.65*	*0.67*
		FAVOUR	0.65	0.73	0.78	0.59	0.66	**0.72**
		PRANK	0.71	0.80	**0.81**	0.47	0.63	0.70
	RLO ≥ 0.5	*P2RANK*	*0.86*	*0.91*	***0.94***	*0.86*	*0.91*	***0.94***
		FAVOUR	0.74	0.83	0.87	0.74	0.84	0.92
		PRANK	0.78	0.89	0.90	0.62	0.81	0.89
	RLO = 1	*P2RANK*	*0.37*	*0.41*	*0.43*	*0.37*	*0.41*	***0.43***
		FAVOUR	0.38	0.43	**0.47**	0.33	0.39	0.42
		PRANK	0.39	0.46	**0.47**	0.28	0.36	0.40
Cryptic Site Pairs	*RRO* ≥ 0.5	*P2RANK*	*0.36*	*0.42*	***0.47***	*0.36*	*0.42*	***0.47***
		FAVOUR	0.29	0.36	0.38	0.25	0.34	0.40
		PRANK	0.28	0.36	0.40	0.20	0.30	0.33
	RLO ≥ 0.5	*P2RANK*	*0.57*	*0.64*	***0.69***	*0.57*	*0.64*	*0.69*
		FAVOUR	0.49	0.60	0.64	0.51	0.64	**0.71**
		PRANK	0.48	0.59	0.64	0.41	0.59	0.64
	RLO = 1	*P2RANK*	*0.12*	*0.14*	*0.15*	*0.12*	*0.14*	***0.15***
		FAVOUR	0.14	0.19	**0.19**	0.10	0.13	0.15
		PRANK	0.14	0.16	0.17	0.07	0.12	0.14
Comparative Models	*RRO* ≥ 0.5	*P2RANK*	*0.27*	*0.32*	*0.36*	*0.27*	*0.32*	***0.36***
		FAVOUR	0.36	0.36	0.41	0.27	0.27	**0.36**
		PRANK	0.36	0.41	**0.45**	0.05	0.05	0.05
	RLO ≥ 0.5	*P2RANK*	*0.55*	*0.64*	***0.68***	*0.55*	*0.64*	***0.68***
		FAVOUR	0.41	0.50	0.59	0.41	0.50	0.64
		PRANK	0.55	0.59	0.64	0.32	0.45	0.50
	RLO = 1	*P2RANK*	*0.14*	*0.18*	*0.18*	*0.14*	*0.18*	***0.18***
		FAVOUR	0.18	0.18	0.18	0.14	0.14	0.14
		PRANK	0.18	0.23	**0.23**	0.14	0.18	**0.18**
Allosteric Sites	*RRO* ≥ 0.5	*P2RANK*	*0.17*	*0.25*	*0.37*	*0.17*	*0.25*	*0.37*
		FAVOUR	0.21	0.31	0.36	0.20	0.31	**0.40**
		PRANK	0.20	0.30	**0.41**	0.12	0.20	0.22
	RLO ≥ 0.5	*P2RANK*	*0.29*	*0.39*	***0.54***	*0.29*	*0.39*	*0.54*
		FAVOUR	0.25	0.37	0.44	0.28	0.43	**0.56**
		PRANK	0.28	0.40	0.52	0.18	0.31	0.36
	RLO = 1	*P2RANK*	*0.11*	*0.16*	*0.20*	*0.11*	*0.16*	***0.20***
		FAVOUR	0.10	0.13	0.16	0.12	0.16	0.20
		PRANK	0.10	0.16	**0.21**	0.06	0.12	0.13

4.1.11 Conclusion

An objective and reliable analysis of a large set of binding site prediction approaches is hampered by several properties which are inherently connected to the nature of small molecule binding sites.

First, binding sites are flexible and their boundaries vary depending on the reference structure. Second, a cavity definition based on the bound ligands (even when using an ensemble of ligands) is restricted to our structural knowledge and cannot be regarded as universally applicable. This can be nicely illustrated for protein-protein interfaces which are shallow and characterised by a large surface area. Nevertheless, the small molecule PPI inhibitors occupy only parts and sub-pockets of these huge interfaces.

Moreover, varying demands on binding site prediction methods exist. While the accuracy of the predicted residues is essential to extract a suitable pocket for binding site comparison tools, most docking approaches accept a single point to derive the cavity which is subsequently used for the pose prediction steps.

Finally, binding sites that are not easily detectable for most methods were highlighted. Besides very small pockets, binding sites that change their conformation upon ligand binding often remain hidden. This hampers the prediction of novel, potentially druggable binding sites.

Despite these challenges, it is possible to derive some guidance for using these approaches after an in-depth analysis of the individual strengths and weaknesses of the multitude of publicly available prediction methodologies. The decision for a suitable method is inevitable for the successful inclusion of binding site prediction in scientific projects such as target identification, drug repurposing, or function prediction. In this study, the hit rates for different methods and datasets were evaluated. The robustness with respect to conformational changes in the pockets was assessed. This is especially interesting for the cryptic site pairs which show higher RMSD values as compared to the ligand-bound and unbound sites of the extended PDBbind core set.

Based on the hit rates, multiple methods, irrespective of the underlying approach (geometry-, energy-, or knowledge-based), show a good performance and can be readily applied. But taking robustness into account, the knowledge-based methods are superior. Consequently, these tools should be preferentially used if the number of known structures is low. Their second strength is the robustness with respect to models derived from comparative modelling. Nevertheless, it is not possible to pinpoint a single method which fulfils all needs. The huge dependence on the dataset used for the evaluation (number of structures, quaternary structures, etc.) leads to different performance rankings of the methods. Whenever the researcher wants to focus on a specific type of site (sites in unbound structures, allosteric sites, sites in comparative models), the choice has to consider the individual limitations of the available methods. There is always a certain number of sites which are difficult to predict as they have special and unique properties.

Even meta approaches often fail to correctly predict sites which are prone to conformational changes or show unique geometric or physicochemical characteristics.

For small sets of structures and restricted structural knowledge, MD simulations and the application of tools such as PocketAnalyzer[PCA] or MDpocket might be appropriate for some of these challenges. Binding site identification should better be regarded as one step in a sequence of methods which enable the reliable prediction of biologically relevant pockets.[37]

Even the establishment of meta approaches is hampered by the difficulties arising when it comes to the scoring of the predicted sites. While it can be stated that all methods are able to correctly detect most of the sites, it is difficult to derive a uniformly applicable ranking measure. The superiority of the method P2RANK in nearly all analyses hints at the importance of this approach and defines a promising future direction for binding site identification. The establishment of new “ligandability” scores based on even larger datasets incorporating multiple types of sites might enable highly reliable pocket predictions.

4.2 Binding Site Comparison

4.2.1 Motivation

Binding site comparison might have become a standard structure-based approach by now as it has been successfully applied for protein function prediction,[52] the establishment of polypharmacology,[351] the identification of novel hit compounds,[352] and other interesting fields of application.[48] Nevertheless, it is still difficult to decide on a tool which can be partially attributed to the large number of publicly available methods. An independent evaluation of the individual strengths and weaknesses of the available binding site comparison tools is still wanting. Most methods were evaluated by the developers using a broad spectrum of highly different datasets. This hinders the choice of a suitable binding site comparison tool for a specific task. Table 4.2.1 gives an overview of the tools which were analysed in more detail herein. Some of the datasets which were used for assessing their individual strengths are also listed. The summary illustrates the huge diversity of the performance assessments, but also the successful application of the approaches in independent studies. Besides examples of method validations on very small sets of protein binding sites, datasets which are not sufficiently diverse were used. Some sets even include protein binding site pairs with unknown similarities. In general, we cannot state that proteins binding to similar ligands necessarily share common cavity properties.[267] The inclusion of binding sites that were shown to be similar by another comparison tool will also lead to an undesirable bias of the dataset. Therefore, a distinct set of state-of-the-art comparison methods was evaluated using novel objective datasets which focus on different quality.

4.2.2 Choice of Binding Site Comparison Tools for Evaluation

Several methods that were successfully applied in independent applications in the field of medicinal chemistry are available as standalone tools for performance assessment. They were analysed herein to elucidate their strengths and limitations. These methods include site comparison methods which are based on fingerprints (FuzCav,[73] KRIPO,[74] PocketMatch,[49] SiteAlign,[77] TIFP[268]), graphs (Cavbase,[64,65] Grim,[268] IsoMIF,[269] ProBiS,[61] SMAP[69]), and grids (VolSite/Shaper[50]), and such that make use of even other, alternative approaches (RAPMAD,[270] SiteEngine,[271] SiteHopper[272]).

Additionally, TM-align[71] was included in this analysis although it is not necessarily a binding site comparison method. It was originally designed to compare proteins based on their overall structure. Nevertheless, the software was successfully applied in various drug discovery scenarios (see Table 4.2.1). Grim and RAPMAD were included in this analysis to evaluate the impact of the underlying data structure (fingerprints vs. graph models for TIFP and Grim and histograms vs. graph models for RAPMAD and Cavbase) on the comparison results. While fingerprint- and histogram-based methods are usually characterised by low computational demands, graph models often describe the binding site features in a more accurate way accompanied by higher runtimes. However, the results presented in Section 4.2.7 show that this basic assumption holds only partially true.

Successfully applied web servers were excluded from the analyses. Due to the number of analysed datasets and their size, an evaluation of these approaches became infeasible. In addition to the tools stated above, the freely available methods CMASA,[81] COFACTOR,[58] and PocketFEATURE[78] could not be used for the benchmark analyses. CMASA is only applicable to internal, already prepared datasets. The same holds true for COFACTOR. Both tools only enable the comparison of the binding site of interest against sets of prepared binding sites provided by the developers. PocketFEATURE was not available from the SimTK web site.

Table 4.2.1. Overview of the binding site comparison methods analysed herein. All methods differ in the way the binding site is represented (binding site modelling) and their quality was assessed using different benchmark datasets. Nevertheless, they share common application domains.

Method	Availability	Benchmark Datasets (Examples)	Application (Note)
Cavbase[64,65]	CCDC	query-based evaluation	protein-ligand interactions,[353] VS,[354] evolutionary relationships,[355] drug repurposing[356]
FuzCav[73]	upon request	set of non-redundant similar and dissimilar binding sites based on SiteAlign comparisons, clustering of ATP and protease binding sites	protein-ligand interactions[357]
Grim[268]	upon request	similar and dissimilar protein-ligand complexes based on a Shaper comparison, rescoring of docking poses	(like TIFP, but graph models are used)
IsoMIF[269] (based on IsoCleft-Finder[358])	download	Kahraman dataset,[265] SOIPPA dataset[70] as modified by Brylinski *et al.* (228 adenine-containing binding sites and 91 true negatives)[359]	drug repurposing[360]
KRIPO[74]	download	fragments which are at least 20 times in the PDB, 10,000 random fragment pairs within 200 diverse PDB ligand fragments (decoy)	off-target prediction[361–363]
PocketMatch[49]	download	similar SCOP classification (family level), sites for identical ligands in multiple protein subunits, binding sites in proteins binding CIT, MTX, MK1, or PGA	function, prediction,[364] poly-pharmacology,[365] evolutionary relationships[366]
ProBiS[61]	download	test sets for binding site detection	function prediction,[367] off-target prediction[368]
RAPMAD[270]	CCDC	dataset of ATP, NAD, and FAD binding sites,[369] dataset of Hoffmann *et al.*[266]	(like Cavbase, but histogram-based)
VolSite/Shaper[50]	upon request	clustering of ATP and protease binding sites	protein-ligand interactions[357]
SiteAlign[77]	upon request	376 non-redundant related binding site pairs (EC classes) from the sc-PDB[10]	protein-ligand interactions[357,370]

Table 4.2.1 *(continued)*. Overview of the binding site comparison methods analysed herein. All methods differ in the way the binding site is represented (binding site modelling) and their quality was assessed using different benchmark datasets. Nevertheless, they share common application domains.

Method	Availability	Benchmark Datasets (Examples)	Application (Note)
SiteEngine[271]	download	query-based searches including adenine-binding proteins, proteases, estradiol-binding proteins, etc. (screening databases of binding sites or complete structures with a known binding site, or databases of binding sites with complete protein structures)	PPIs[371]
Sitehopper[272]	OpenEye	query-based searches of sc-PDB entries including therapeutically relevant targets	evolutionary relationships[372]
SMAP[69] (based on SOIPPA[70])	download	247 non-redundant protein chains binding an adenine-containing ligand and 101 non-redundant protein chains believed not to bind an adenine-containing molecule	poly-pharmacology,[373] drug repurposing[54,374]
TIFP[268]	upon request	similar and dissimilar protein-ligand complexes based on a Shaper comparison, rescoring of docking poses	VS[375]
TM-align[71]	download	200 non-homologous PDB proteins (comparison to other protein structure alignment algorithms)	drug repurposing[376]

4.2.3 Binding Site Comparison Methodologies

The outcome of the binding site comparison tool depends on the description of the binding site. In the methods discussed herein, binding sites are modelled in three ways: (a) based on the residues lining the binding site of the ligand, (b) based on the ligand atoms interacting with the binding site residues, and finally (c) based on the interaction between the bound ligand and the binding site residues. In general, only the first method does not rely on bound ligands, while tools that compare interactions often require a ligand occupying the binding site. Therefore, methods considering protein properties irrespective of the ligand are applicable to proposed novel binding sites. These binding sites can be derived from previous pocket predictions by various binding site detection tools as illustrated in Section 4.1. Additionally, novel binding sites in apo structures of proteins can also be identified utilising binding site comparison, as for example shown for ProBiS[61] and VolSite/Shaper.[50]

Besides physicochemical properties, there are numerous further binding site descriptors that were successfully applied to compare binding sites, e.g., the shape of the binding site, geometric potentials, and SSEs. The methods analysed herein describe the binding sites with the help of the binding site residues, by binding site surface patches, or by means of protein-ligand interaction patterns. All methods that were investigated for benchmark purposes are presented in Figure 4.2.1. It depicts the methods, the way the binding sites are modelled, and the data structures which are subsequently used for the similarity searches.

In the following paragraphs, the general approaches underlying the methods under investigation are introduced.

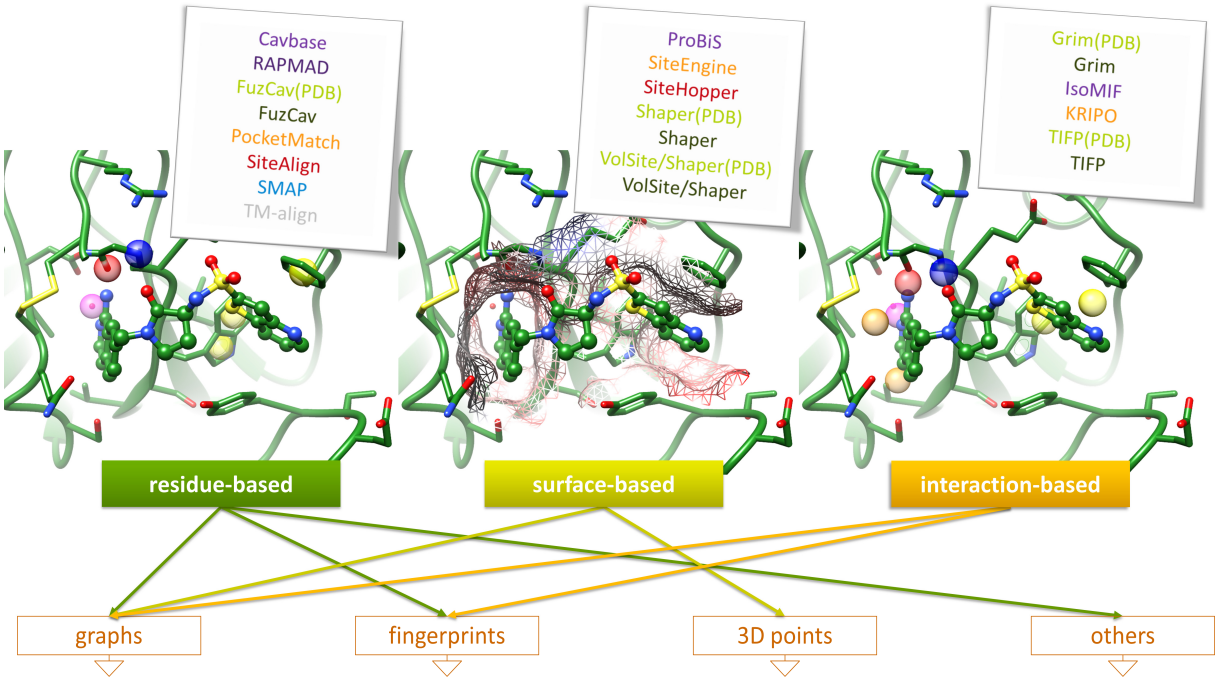


Figure 4.2.1. Ways of binding site modelling that are applied by binding site comparison methods. The chosen example shows the binding site of human coagulation factor Xa (PDB-ID 1f0r, chain A) in complex with the nanomolar inhibitor RPR208815 in three binding site representation schemes. The arrows indicate the corresponding underlying data structures for the comparison. The binding site representations were generated with UCSF Chimera.[228] The font colours of the methods are used throughout this section to indicate the methods' performance for the benchmark datasets.

Residue-Based Binding Site Comparison Methods

Cavbase and RAPMAD[64,65,270] The properties of the binding site residues are represented by means of pseudo-centres in both methods. The physicochemical properties of the binding site residue atoms (detected by an implementation of the pocket detection approach LIGSITE[273]) are assigned with the following types: hydrogen bond acceptor or donor, mixed hydrogen bond acceptor/donor, centre of an aromatic ring, π electron centre, aliphatic moiety, and metal ion. In Cavbase, these pseudo-centres are matched to identify suitable superposition possibilities. To this end, the following step transforms the pseudo-centres into undirected graphs. A search for the MCS is performed to obtain binding site alignments. The similarity of the final match is calculated using the number of the matching surface points within a distance of 1 Å. The similarity score includes the mutual overlap of the points from two matching surface patches, the RMSD, and the number of matching pseudo-centre pairs. In RAPMAD (RApid Pocket MAtching using Distances), the pseudo-centres are divided into seven property subsets and the centroid and pseudo-centre closest to this centroid are determined per set. The Euclidean distances of these two reference points to all pseudo-centres enables the generation of distance histograms (with a bin size of 0.4 Å) for the comparison. The tool calculates a distance measure for the compared cavities based on the histogram distributions (Jensen-Shannon divergence). The outcome of both methods was shown to be comparable with respect to their classification performance.

FuzCav[77] The Cα atom coordinates of the binding site residues are labelled with six pharmacophoric features (hydrogen bond acceptor and donor, positive and negative ionisable, aromatic, aliphatic). These features and their spatial relationships are summarised by a fingerprint representation which encodes all unique pharmacophore triplets and the binned Cα atom distances. The fingerprint for 4,833 possible triplets is used for the comparison. The similarity score is calculated by dividing the number of common fingerprint strings by the minimum number of non-zero fingerprint strings of the compared binding sites.

PocketMatch[49] The binding site residues are grouped by their physicochemical properties (0: Ala, Val, Ile, Leu, Gly, Pro, Met; 1: Lys, Arg, His; 2: Asp, Asn, Glu, Gln; 3: Tyr, Phe, Trp; 4: Cys, Ser, Thr) and are represented by three types of points (Cα, Cβ, and Ccentroid, i.e., the geometric centre of the side-chain atoms). The binning of all pairs of point distances (group-type pairs and point-type pairs) enables the comparison of pairs of sorted distance sequences. The recommended scoring scheme divides the number of distance pair matches by the total number of distance pairs of the larger cavity ($PMScore_{max}$). The $PMScore_{min}$ takes the number of distance pairs of the smaller cavity into account.

SiteAlign[77] Topological characteristics (distance Cβ - sphere centre, side-chain orientation, size) and physicochemical properties (hydrogen bond acceptor or donor, charge, aromatic, aliphatic) are used to describe the binding site. They are projected on a discretised 80 triangle sphere with its origin in the cavity centre by deriving a geometric vector from the Cα atom of each binding site residue to the sphere centre. The eight descriptors are mapped on the respective triangle to generate the binding site fingerprint. The alignment seeks to optimise the number of possible similarities between both sites. For the final scoring, the normalised similarities per triangle and descriptor are calculated and averaged per triangle. Four distances (d1, d2, d3, d4) are reported by subtracting the following similarities from 1: the triangle similarity sum which is divided by the number of the non-zero triangles for one of both sites (s1), the number of the non-zero triangles for both pockets (s2), the number of the non-zero triangles which lead to local distances smaller than or equal to 0.2 for both sites (s3), and the number of the non-zero triangles with polar character for both sites (s4).

SMAP[69] The previously developed method SOIPPA (Sequence Order Independent Profile-Profile Alignments)[70] constitutes the basis for SMAP. A mesh surface is generated by a Delaunay triangulation of the Cα atoms of the binding site. Each Cα atom is assigned a normal vector that is perpendicular to the surface. Moreover, a PSI-BLAST[244] search is conducted to derive the PSSM for the binding site residues and assign a probability distribution to each Cα atom. The annotated ligand binding site is finally represented as a graph which is based on the regular tessellation. The McLachlan[377] similarity matrix describes the physicochemical and evolutionary relationship of the matched cavities. The angle and distance between two matched atoms are used as the geometric quality criteria. The resulting score (RawScore) is returned, together with the RMSD, the template and query coverage, and the Tanimoto coefficient that describes the ratio of the overlapping pocket volume (intersection) and the union pocket volumes of both sites.

In contrast to SOIPPA, the statistical model implemented in SMAP assists in assessing the statistical significance of the match by the calculation of the p-values which are based on an extreme value distribution. The method is applicable to detect local binding site similarities as well as those between cavities and complete proteins.

TM-align[71] The tool was originally developed to superpose complete protein structures but was also successfully applied to site comparisons. First, a sequence- and SSE-based alignment is generated. The subsequent superposition is based on the Cα coordinates of the aligned residues. The similarity measure (TM-score) does not consider the physicochemical nature of the matched residues. It represents the sum of distances between the aligned residues after the optimal spatial superposition. It is noteworthy that the algorithm is sequence order dependent, i.e., the discontinuous alignment of the residues is not possible.

Surface-Based Binding Site Comparison Methods

ProBiS[61] The method compares binding sites using graphs. The SAS atoms of the cavity are annotated by the nature of their underlying functional group (hydrogen bond acceptor or donor, mixed hydrogen bond acceptor/donor, aromatic, and aliphatic). A labelled graph comprising the atoms as vertices is used for the comparison. Subgraphs are derived by connecting the vertices that are within a 15 Å distance for the use of complete protein structures. Filtering steps enable fast comparisons. The tool returns the expectation value (e-value) according to the Karlin-Altschul equation,[378] the surface vector angle between the two normal vectors of the matched protein surface patches (SVA), the RMSD of the aligned surface vertices, the z-score (calculated from the number of matched vertices, the e-value, and the RMSD), and the Alignment Score that is composed of all terms. The method enables the comparison of binding sites to both binding sites and proteins, and proteins to proteins to identify local similarities.

SiteEngine[62] The physicochemical site properties are determined similarly to the Cavbase approach. The binding site residue atoms (4 Å around the binding partner) are grouped according to their potential ligand interactions. A Connolly surface of the binding site is constructed and the pseudo-centres with at least one surface-exposed atom are retained. Pseudo-centre triplets are generated and encoded with the respective side lengths and a physicochemical index. The comparison via hashing and the elaborate match analysis and rescoring of possible transformations lead to the identification of binding site similarities. All identified possible matches per pair are filtered by applying an additional graph matching procedure to identify the maximum weight match. The results can be ranked according to the LowResolutionScore, the OverallSurfaceScore, the DistanceScore, the CurvatureScore, or the TotalScore that is calculated by combining the first four scoring metrics. The tool allows for the comparison of pockets as well as protein structures and pockets.

SiteHopper[272] The ligand-occupied cavity is represented as a 3D shape that is “coloured” by the chemical properties of the underlying binding site residues. This is achieved with the help of the Shape[379] and Spicoli[380] toolkits of OpenEye. After the alignment of the binding site patches, the maximum overlap of the atom-centred Gaussian functions is calculated.

The final so-called PatchScore combines a measure for physicochemical similarity (ColorTanimoto) and a measure to estimate the surface shape similarity (ShapeTanimoto). The PatchScore is calculated by summing up the ShapeTanimoto and the threefold ColorTanimoto. The software enables comparisons of protein pockets but does not support the comparison of pockets against protein structures or vice versa. In a recent publication, it was shown that the tool is also applicable to predicted binding sites as defined by fpocket.[196]

VolSite/Shaper[50] VolSite is used to generate the binding site representation that is used for comparisons with Shaper. A cube is placed at the ligand's centre of mass and filled with a uniform grid. Predicted binding sites can also be used for the binding site definition. "IN" and "OUT" properties are assigned to the grid cells depending on the presence or absence of protein atoms and their buriedness. The resulting site points are assigned a physicochemical property (hydrogen bond acceptor or donor, mixed hydrogen bond acceptor/donor, hydrophobic, aromatic, positive or negative ionisable). The VolSite attributes are additionally used for druggability prediction by an SVM-based algorithm. Druggable annotated binding sites are subsequently used for the cavity comparisons with Shaper. This tool is designed for pharmacophore-annotated cavity shape comparisons. Various similarity measures are provided: the color and fit Tanimoto and their sum (combo), the color and fit Tversky index and their sum (combo) with a higher weight on the reference cavity (RefTversky) or the superposed cavity (FitTversky).

Interaction-Based Binding Site Comparison Methods

IsoMIF[269] The method is based on the previously published tool IsoCleft Finder.[358] The software compares MIFs which describe the putative interactions of binding site residues with ligand atoms. Pockets are detected by the GetCleft[197] algorithm (see Section 4.1.2). Moreover, cavities can be defined by the bound ligand. A grid with a 0.5 Å spacing and a distance of 2.5 to 4 Å to any protein non-hydrogen atom is then generated. Each grid point is labelled with the interaction potential energy for six chemical probes (hydrophobic, aromatic, hydrogen bond acceptor or donor, positively or negatively charged). The protein atom types are assigned according to the original SYBYL atom types[381] and an additional "n.his" atom type for the histidine nitrogen atoms that can occur in different protonation states (hydrogen bond acceptor and donor).

Based on a given probe interaction energy threshold, probes at the grid points are neglected or retained. For hydrogen bond donor and acceptor probes as well as aromatic interactions, additional threshold angles were introduced. The final MIF is the set of interaction vectors at all grid positions. A graph-matching algorithm is used to identify similarities of the MIFs of two binding sites. The similarity score is calculated as Tanimoto similarity considering the number of common probes and the number of energetically significant probes for each of the two binding sites. Additionally, the RMSD value is given. The resulting match is additionally annotated by three further similarity scores (taniM, taniMW, and taniNormNodes) which are not described in the literature.

KRIPO[74] The binding sites are extracted and described based on bound ligand fragments or ligands in KRIPO (Key Representation of Interactions in Pockets). All residues within a 6 Å radius of the ligand or ligand fragment atoms are defined as the cavity residues. Intermolecular interaction features (hydrogen bond acceptors or donors, aromatic T- and π-interactions, hydrophobic interactions, salt bridges) are described by placing them at distinct geometric positions with respect to the underlying residue according to predefined rules (e.g., hydrogen bond donor vectors are placed at a distinct position along the vectors defined by the N-H and O-H bonds). Distances between interaction features are binned using a non-linear binning scheme and 3-point pharmacophore fingerprints are generated. Each triplet is encoded by the features and the binned distances. The feature opposite the shortest distance is prioritised to provide unique codes. An additional fuzziness was introduced by setting more than one bit in the fingerprint for one pharmacophore triplet, i.e., three distances are encoded for each of the three feature distances. This leads to 27 bits which represent one triplet. A modified Tanimoto coefficient according to Fligner *et al.*[382] was introduced as the similarity measure to reduce the size bias.

TIFP and Grim[268] Protein and ligand atoms are assigned seven physicochemical properties (hydrogen bond acceptor or donor, hydrophobic, aromatic, positive or negative ionisable, metal) for comparisons with TIFP (Interaction Fingerprint Triplets). The detection of the interactions is based on geometric rules including distance and angle thresholds. The interactions are: hydrophobic (hydrophobic - hydrophobic), face-to-face and edge-to-face (aromatic - aromatic), hydrogen bonds ((weak) donor - (weak) acceptor), ionic (anion - cation), π-cation (cation - aromatic), and metal coordination (metal - acceptor). The overall interaction pattern comprises the two interacting atoms of the protein (InterProt) and the ligand (InterLig) as well as the interaction pseudo-atoms (the geometric centre of the two interacting atoms). The user can decide on a suitable comparison mode. Interaction triplets are derived based on the interaction properties and (binned) distances. The derived interaction fingerprint triplets are compared. The similarities can be evaluated by different fingerprint similarity measures. Six fingerprint similarity measures are provided to rank the binding site pairs: the Tanimoto similarity, the Hamming distance, two Tversky indices (RefTversky, FitTversky), the Dice index, and the Soergel similarity.

In contrast, Grim (Graph interaction matching) uses a graph matching procedure for the comparison of interaction pseudo-atoms. The relative frequencies of the interaction types are taken into account by weight assignments which were derived from the frequencies of the observed interactions in the sc-PDB.[10] The final equation to calculate the similarity score Grscore (Graph-alignment score) was derived from the empirical analysis of 900 similar and 900 dissimilar protein-ligand complexes. It results from different contributions of the number of matched InterLig interaction pseudo-atoms, the number of matched centred interaction pseudo-atoms, the sum of the pair weights in the clique divided by the sum of all possible pair weights, the number of matched InterProt interaction pseudo-atoms, the RMSD of the clique, and the absolute value of the difference in the number of interaction pseudo-atoms between the query and the reference site.

4.2.4 Benchmark Datasets

The benchmark datasets were designed to address typical challenges of binding site comparison and are summarised in Table 4.2.2. First, the sensitivity of the selected tools towards the binding site definition was evaluated. Different ligands can address various regions of the binding site of interest (sub-pockets). Therefore, the binding site definition often depends on bound ligands. Similar subparts of pockets are more difficult to match, and the scoring scheme must be optimised for this scenario to enrich binding site pairs with only local similarities in the ranking list (dataset 1). Second, the performance of the tools with respect to binding site flexibility was assessed. This task can easily be achieved by extracting the individual protein models of ligand-bound solution NMR structures that provide an ensemble of structures for one protein (dataset 2). Next, the scoring measures and their performance on structures with identical sequences and various artificially created variants of this structure with one, two, three, four, and five substitutions per binding site (decoy structures) were analysed. To this end, substitution variants were artificially generated by randomly picking surface-exposed binding site residues and substituting them by residues with different physicochemical or shape properties (datasets 3 and 4). A comparison between the original structures and their variants should lead to an enrichment of the original, unmodified site pairs.

Additionally, the dataset of Kahraman *et al.*,[265] which was often used to evaluate whether a tool is able to differentiate between binding sites that bind to different ligands, was prepared to test the methods' ability to identify functionally similar sites (dataset 6). With respect to this dataset, it is worthwhile mentioning that it was designed to evaluate whether the binding pocket shape is related to the shape of the binding ligand. The authors state that the variability of binding site shape cannot solely be explained by the conformational variability of the ligand. Additionally, it is not stated whether there are known similarities between the binding sites in complex with similar ligands. In contrast to this dataset, the dataset of Barelier *et al.*[267] comprises 62 pairs of unrelated proteins binding to similar ligands. In short, it is composed of 19 site pairs with interaction-based similarities, whereas the remaining pairs do not show any obvious binding site similarities (dataset 7). The ability of the methods to enrich the similar site pairs as compared to the dissimilar ones was tested.

To establish an alternative for these two already published datasets, all ligand conformations as stored in the sc-PDB[10] were retrieved and a ROCS-based ligand similarity search in an all-against-all manner was performed. Ligand pairs that showed a high TanimotoCombo and which bind to proteins with a low overall similarity were used as the similar site pairs. Unrelated protein structures with dissimilar ligands were declared to be dissimilar cavity pairs. The basic idea was to construct a dataset which circumvents some of the problems observed in the study of unrelated proteins binding to identical ligands. First, this dataset was enlarged by including not only identical but also similar ligands. Second, it was possible to exclude the influence of ligand flexibility on the dataset. Large and very flexible ligands have various possibilities to interact with the target. In contrast, the inclusion of shape and physicochemical similarity as provided by a 3D similarity analysis leads to an unbiased benchmark dataset.

Table 4.2.2. Summary of the benchmark datasets used to evaluate binding site comparison methods. Mean, standard deviation (±), minimum, and maximum are given for the Resolution, the R_{work}, and the average overall G-Factor of all dataset structures. The average overall G-factors were calculated using PROCHECK[255] and PROCHECK-NMR[256] to enable the quality comparison between all dataset including dataset 2.

Quality Criterion	**Number of Comparisons (Similar / Dissimilar Pairs)**	**Quality Estimates**		
		Resolution [Å]	**R_{work}**	**Average Overall G-Factor**
Dataset 1 (Structures with identical sequences)				
sensitivity with respect to the binding site definition, score range for similar and dissimilar site pairs	13,430 / 92,846 (12 groups of similar sites)	1.79 ± 0.37, 0.8, 2.71	0.17 ± 0.03, 0.09, 0.26	0.02 ± 0.23, -1.27, 0.6
Dataset 1.2 *(Structures with identical sequences and similar ligands)*				
impact of ligand diversity on binding site comparison	*241 / 1,784*	*1.73 ± 0.37, 0.92, 2.5*	*0.17 ± 0.03, 0.1, 0.23*	*0.02 ± 0.22, -0.57, 0.6*
Dataset 2 (NMR structures)				
sensitivity with respect to binding site flexibility	7,729 / 100,512 (17 ensembles)	-	-	-0.28 ± 0.71, -2.8, 0.21
Dataset 3 (Decoy set 1)				
differentiation between binding sites with different physicochemical properties	13,430 / 67,150 (complete dataset) 13,430 / 13,430 (5 residue variants)	-	-	-
Dataset 4 (Decoy set 2)				
differentiation between binding sites with different physicochemical and shape properties	13,430 / 67,150 (complete dataset) 13,430 / 13,430 (5 residue variants)	-	-	-
Dataset 5 (The ROCS dataset)				
identification of similar interaction patterns for similar ligands	15,339 / 56,179	2.18 ± 0.42, 0.8, 3.6	0.19 ± 0.03, 0.1, 0.35	0.12 ± 0.20, -3.19, 1.35
Dataset 6 (The Kahraman dataset)				
classification of proteins binding to identical ligands and cofactors	920 / 5,480	2.02 ± 0.37, 0.88, 2.9	0.20 ± 0.03, 0.09, 0.27	0.17 ± 0.23, -0.56, 0.47
Dataset 6.2 *(The complete Kahraman dataset)*				
original dataset	*1,320 / 8,680*	*2.02 ± 0.4, 0.88, 2.9*	*0.20 ± 0.03, 0.09, 0.27*	*0.16 ± 0.22, -0.56, 0.47*
Datasets 7 (The Barelier dataset) and *7.2 (The Barelier dataset with cofactors)*				
identification of distant relationships between sites with identical ligands "observing" a similar environment	19 / 43	2.16 ± 0.44, 0.93, 3.1	0.20 ± 0.03, 0.10, 0.25	0.12 ± 0.23, -1.46, 0.53
Dataset 8 (The dataset of successful applications)				
recovery of binding site similarities within a set of diverse proteins	115 / 56,284 (49 query structures)	1.98 ± 0.43, 0.8, 3.25	0.19 ± 0.03, 0.10, 0.28	0.13 ± 0.21, -2.8, 1.35

Finally, a set of binding site pairs whose similarities were correctly identified by at least one binding site comparison tool, as summarised in a perspective on the impact of binding site similarity on medicinal chemistry[48], was compiled. These binding sites were combined with a diverse set of binding sites retrieved from the sc-PDB[10] (dataset 8). It was assessed whether the different tools are able to retrieve the known similar pairs within the high-scoring hits. Maybe, this benchmark set contains the most interesting protein pairs as their similarity was shown to be useful in different projects with a focus on structure-based design.

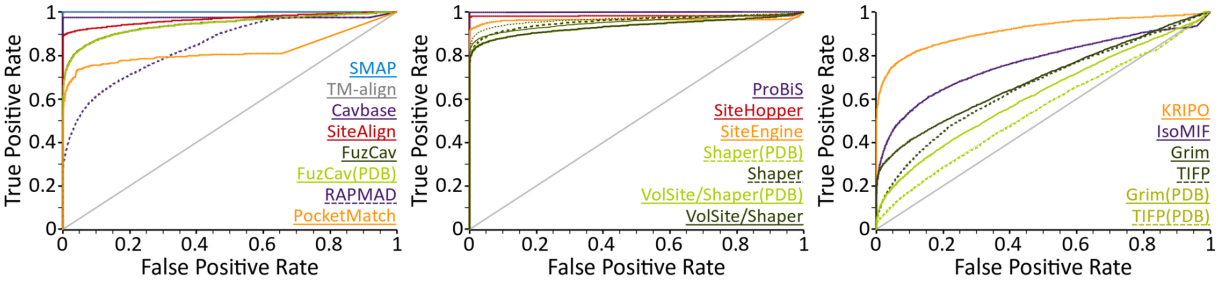
As all the datasets of this study include “active” (i.e., similar) and “inactive” (i.e., dissimilar) pairs, ROC curves could be obtained from all analyses. The EFs at different percentages of the screened dataset were additionally calculated. They will be discussed in the following paragraphs. Tables A2.6, A2.7, A2.10, A2.12, A2.13, A2.16, A2.18, A2.19, A2.22, A2.23, and A2.26 in Appendix 2 provide detailed summaries of all benchmark results. The significance of the AUC values, their confidence intervals (CIs) and the differences between individual AUC values for different tools are provided in Tables A2.8, A2.9, A2.11, A2.14, A2.15, A2.17, A2.20, A2.21, A2.24, A2.25, and A2.27 in Appendix 2.

4.2.5 Benchmark Studies

Structures with Identical Sequences

The first dataset consists of 12 groups of structures with identical sequences whose binding sites are complexed with ligands with different topological and physicochemical characteristics. It was used to evaluate the sensitivity of the methods with respect to the binding site definition. The molecular representations of the respective binding site alignments and the mean pairwise Tanimoto coefficients of the ligands can be found in Figure A2.1 in Appendix 2. Tools which cannot enrich similar binding sites accommodating different ligands in the highest ranked similarity pairs should not be applied for drug repurposing projects or off-target predictions. The resulting ROC curves and EFs are depicted in Figure 4.2.2. Most of the differences between all tools are significant (see Table A2.8) except for ProBiS and SMAP which perform identically for this benchmark set. The impact of the binding site file type (PDB and MOL2 files) on FuzCav, Shaper, and VolSite/Shaper comparisons is negligible, but is high for comparisons with TIFP and Grim.

The plots of the EFs for the dataset reveal that nearly all methods are characterised by a reasonably high early enrichment. Similar sites are ranked highest within a list of site pairs which are sorted by their similarity score. However, nearly all fingerprint-, histogram-, or sorted list-based methods with relatively low computational demands are characterised by a low sensitivity for this dataset. Overall, FuzCav, PocketMatch, and RAPMAD are highly dependent on the chemical nature and the size of the binding site-defining ligand. This trend is recognisable when comparing the related tools Cavbase and RAPMAD or Grim and TIFP. While RAPMAD and TIFP perform fast comparisons, their overall accuracy is lower than that of their graph-based counterparts. Nevertheless, the overall early enrichment is promising in all cases underlining the general applicability of all methods analysed herein.



	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	7.9	7.9	7.9	7.9	7.9	7.9	7.9
FuzCav	7.9	7.9	7.9	7.9	7.9	7.9	7.9
FuzCav (PDB)	7.9	7.9	7.9	7.9	7.9	7.9	7.9
PocketMatch	7.6	7.8	7.9	7.9	7.9	7.9	7.9
RAPMAD	7.9	7.9	7.9	7.9	7.8	7.4	6.9
SiteAlign	7.9	7.9	7.9	7.9	7.9	7.9	7.9
SMAP	7.9	7.9	7.9	7.9	7.9	7.9	7.9
TM-align	7.9	7.9	7.9	7.9	7.9	7.9	7.9

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
ProBiS	7.9	7.9	7.9	7.9	7.9	7.9	7.9
Shaper	7.9	7.9	7.9	7.9	7.9	7.9	7.9
Shaper (PDB)	7.9	7.9	7.9	7.9	7.9	7.9	7.9
SiteEngine	7.9	7.9	7.9	7.9	7.9	7.9	7.9
SiteHopper	7.9	7.9	7.9	7.9	7.9	7.9	7.9
VolSite/Shaper	7.9	7.9	7.9	7.9	7.9	7.9	7.9
VolSite/Shaper (PDB)	7.9	7.9	7.9	7.9	7.9	7.9	7.9

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Grim	7.9	7.9	7.8	7.8	7.4	6.5	5.7
Grim (PDB)	6.3	6.8	5.5	4.5	4.0	3.6	3.3
IsoMIF	7.9	7.9	7.8	7.5	7.1	6.6	6.1
KRIPO	7.9	7.9	7.9	7.9	7.9	7.9	7.9
TIFP	7.9	7.2	5.4	4.4	4.0	3.7	3.4
TIFP (PDB)	1.6	5.5	4.1	2.9	2.4	2.2	2.0

Figure 4.2.2. Evaluation of different binding site comparison tools with respect to dataset 1. Top: ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is coloured according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. Thin lines represent the resulting ROC curve for the scoring scheme that yielded the highest AUC. For the surface-based methods, the Tanimoto (color) and ColorTanimoto for Shaper or VolSite/Shaper and SiteHopper led to the highest AUC, respectively. The use of the Tanimoto coefficient as the similarity measure led to the highest AUC for TIFP(PDB). Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

The effect of ligand differences on the scoring of similar pocket pairs is most pronounced for the comparison tools that rely on distinct protein-ligand interaction patterns. This can be explained by the fact that identical proteins interact differently with their ligands. Figure 4.2.3 illustrates the described challenge for one example. Interaction-based tools should be used in combination with a large set of available protein-ligand structures. If the number of available complex structures for the POI is especially low, it might even be a promising approach to apply molecular docking studies for further known ligands and use their resulting binding poses in the site for the subsequent comparison, in addition to the experimentally solved complex structures. The exceptions are KRIPO and IsoMIF. In the case of IsoMIF, ligand-defined binding sites without information on the interactions to the bound ligands are used. The original binding site definition is therefore only partially influenced by the ligand. Similarly, KRIPO interaction fingerprints are derived based on the residues of the pocket.

The performance of interaction-based methods increases for a reduced set of structures with identical sequences (see Figure A2.2 in Appendix 2). This set includes only 45 binding sites whose ligands are characterised by pairwise Tanimoto coefficients of at least 0.6 (see Table A2.28). In consequence, the interaction patterns of the ligands are more similar. The exclusion of dissimilar ligands leads to improvements for nearly all methods analysed. However, this increase in performance is small for the residue- and surface-based methods as most of them performed nearly perfectly for the original dataset. These methods should be preferred for drug repurposing, polypharmacology analyses, or off-target predictions as they do not depend on the chemical nature and the interactions of the bound ligands.

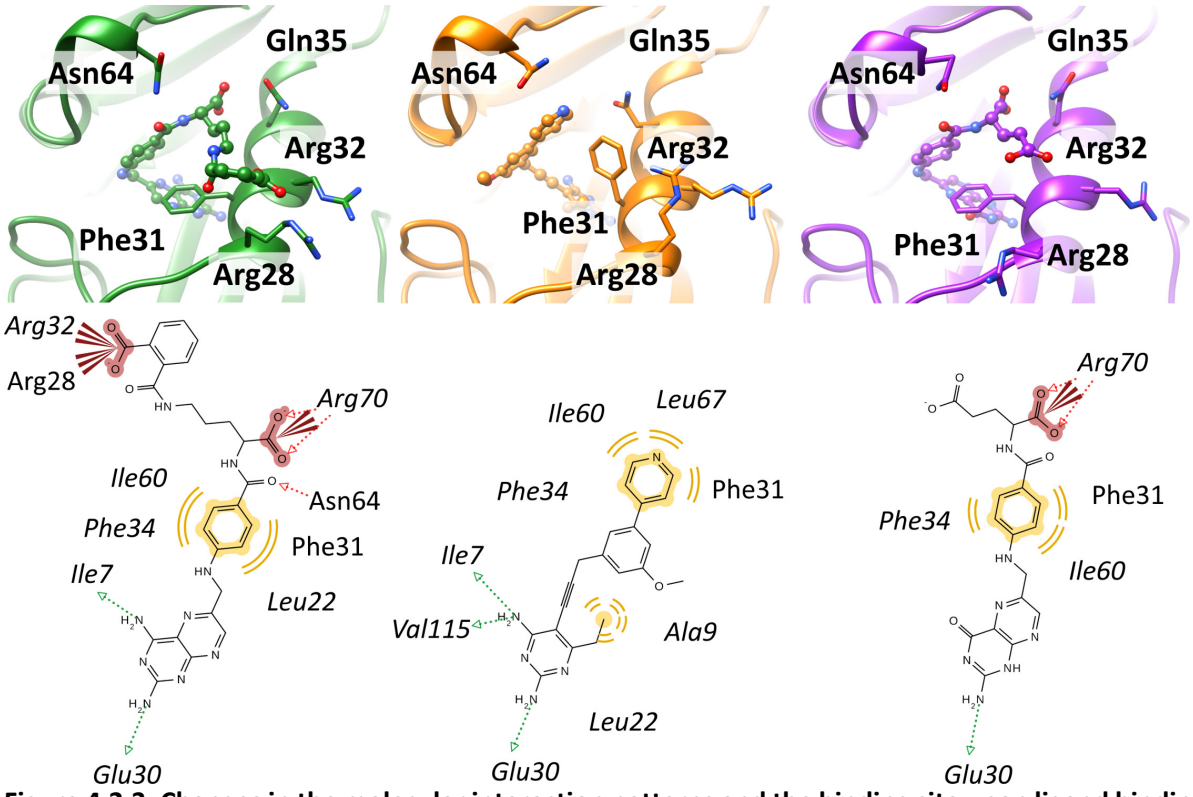


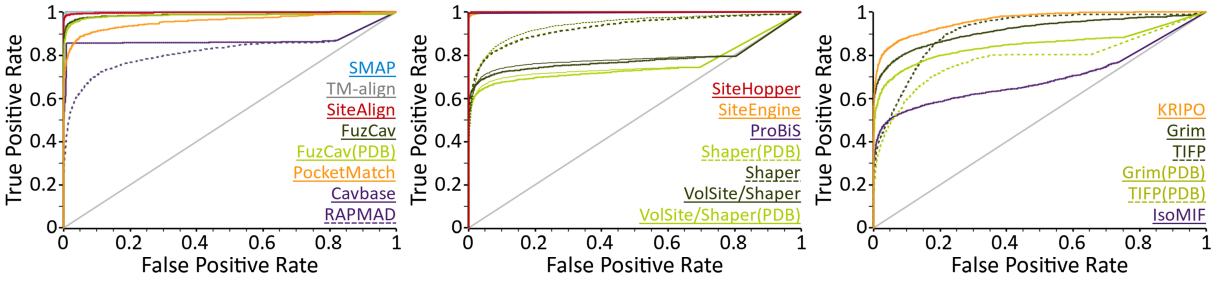
Figure 4.2.3. Changes in the molecular interaction patterns and the binding site upon ligand binding. Differences in the molecular interaction patterns of dihydrofolate reductase ligands upon binding are presented (left: PDB-ID 1ohk chain A, centre: PDB-ID 4kd7 chain A, right: PDB-ID 1drf chain A). Top: Representations of the binding site structures. The figures were generated with UCSF Chimera.[228] Bottom: Schematic view of the interactions between protein and ligand. The figures were generated with LigandScout 4.0.[383]

An additional analysis focused on the ability of the scoring schemes to discriminate between similar and dissimilar site pairs. As recently shown, the application of different scoring schemes might considerably influence the results.[384] Figures A2.3 to A2.5 show boxplots illustrating the distribution of scores for “active” (similar) and “inactive” (dissimilar) pairs. These figures provide the score ranges for similar binding sites for the examined tools. With respect to the residue-based methods, we see a clear score-dependent distinction between similar and dissimilar site pairs for Cavbase, SiteAlign, and TM-align. Considering the surface-based tools, the ProBiS Alignment Score and the PatchScore of SiteHopper are best suited for a clear differentiation between similar and dissimilar pairs. The basic benefit of a large gap between the scores of both pairs is the fact that more distinct, but similar binding sites have a higher chance of being scored higher than the dissimilar pairs. In comparison, the scoring schemes of all interaction-based methods do not provide such discrete score ranges for similar and dissimilar cavities. Within this selection of tools, IsoMIF and KRIPO performed best. A possible explanation was given above. The chemically different ligands located at the same binding site are involved in different interactions which leads to very different interaction patterns. Nevertheless, the abstract interaction-based binding site description can lead to interesting findings, e.g., unrelated proteins binding to similar ligands (which is discussed in more detail for dataset 6). Finally, it is important to note that although boxplots are useful to estimate meaningful score ranges, fixed score thresholds might lead to misinterpretations concerning similarities and should therefore be avoided (see also Section 4.2.6).

In conclusion, the application of interaction-based methods to find similarities between proteins binding to chemically different ligands (e.g., in drug repurposing studies or for off-target predictions) can only be successful as part of an elaborate workflow. The use of binding site prediction, molecular docking, and a detailed visual inspection of the identified similarities are indispensable for TIFP, Grim, IsoMIF, and KRIPO. The use of "composite ligands" might also improve the results, i.e., if the common interaction patterns of multiple ligands are used. In contrast, residue- and surface-based methods are the best choice for the elucidation of relationships between related sites in complex with different ligands.

NMR Structures

A detailed examination of Figure 4.2.3 reveals that not only huge differences in interaction patterns can lead to low similarity scores for similar binding sites, but also different binding site conformations. Induced fit and conformational selection effects lead to changes in side-chain orientations and sometimes also to changes in the backbone geometry of a protein. Such observations prompted the design of a second dataset which can shed light on the degree to which conformational changes influence the outcome of the binding site comparison methods. Solution NMR structures provide access to a sampled conformational ensemble of protein structures. Therefore, the individual model structures of the sc-PDB-derived NMR complexes were used to evaluate whether the tools can recover similar binding sites in different conformations. In contrast to dataset 1, the structures within the groups of identical binding sites are characterised by higher pairwise RMSD values (see Table A2.29). Consequently, the basic challenge of this dataset was binding site flexibility. The detailed analysis of all tools is depicted in Figure 4.2.4.



	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	14.0	14.0	14.0	14.0	14.0	13.1	12.3
FuzCav	14.0	14.0	14.0	14.0	14.0	14.0	14.0
FuzCav (PDB)	14.0	14.0	14.0	14.0	14.0	14.0	13.9
PocketMatch	14.0	14.0	14.0	14.0	14.0	13.9	13.5
RAPMAD	14.0	14.0	13.9	12.9	11.7	10.7	9.9
SiteAlign	14.0	14.0	14.0	14.0	14.0	14.0	14.0
SMAP	14.0	14.0	14.0	14.0	14.0	14.0	14.0
TM-align	14.0	14.0	14.0	14.0	14.0	14.0	14.0

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
ProBiS	14.0	14.0	14.0	14.0	14.0	14.0	14.0
Shaper	14.0	14.0	14.0	14.0	13.8	13.2	12.0
Shaper (PDB)	14.0	14.0	14.0	14.0	13.8	13.2	12.2
SiteEngine	14.0	14.0	14.0	14.0	14.0	14.0	14.0
SiteHopper	14.0	14.0	14.0	14.0	14.0	14.0	14.0
VolSite/Shaper	14.0	14.0	14.0	14.0	14.0	13.6	12.3
VolSite/Shaper (PDB)	14.0	14.0	14.0	14.0	13.9	13.2	11.6

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Grim	14.0	14.0	14.0	14.0	14.0	13.8	12.6
Grim (PDB)	14.0	13.8	13.9	13.5	13.4	12.6	11.3
IsoMIF	14.0	13.9	13.9	13.3	11.8	10.2	8.9
KRIPO	14.0	14.0	14.0	14.0	13.9	13.6	12.9
TIFP	14.0	13.6	13.1	11.7	10.8	9.6	8.4
TIFP (PDB)	6.0	12.4	12.6	11.2	9.8	8.8	7.8

Figure 4.2.4. Evaluation of different binding site comparison tools with respect to dataset 2. Top: ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is coloured according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. The highest AUC was obtained for SiteAlign when using the distance measure d1. All Shaper comparisons led to higher AUCs for the score Tanimoto (color). SiteEngine results slightly improved with respect to the AUC for the DistanceScore. Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

The impact of binding site flexibility on the performance of the tools is lower as compared to that of different binding site definitions. However, there are some methods which do not tolerate different binding site conformations. Shaper and IsoMIF are characterised by a poor performance for this dataset which is significantly lower than that for dataset 1.

The significant difference between the results of using Shaper or Shaper in combination with VolSite was investigated in more detail. The results point towards some major challenges with respect to druggability prediction.[8] VolSite initially calculates the druggability score for the binding sites of interest. If this score drops below a predefined threshold, the binding site is regarded as non-druggable. In this case, no cavities are provided and can therefore not be used for comparison purposes. Intriguingly, some models of one solution NMR ensemble were predicted to be druggable, while others were not. To independently analyse this in more detail, DogSiteScorer[191] was used to calculate the properties of the NMR model-derived binding sites. Figure 4.2.5 shows the high deviations of the number of binding site acceptor and donor atoms, aromatic atoms, and large differences in hydrophobicity for models within the NMR ensembles of dataset 2. It illustrates the high fluctuations in the binding site properties due to the flexibility of the protein (see Table A2.30 for further details) and highlights that druggability predictions should be taken with care and performed for multiple representatives of one protein structure to obtain reliable results.

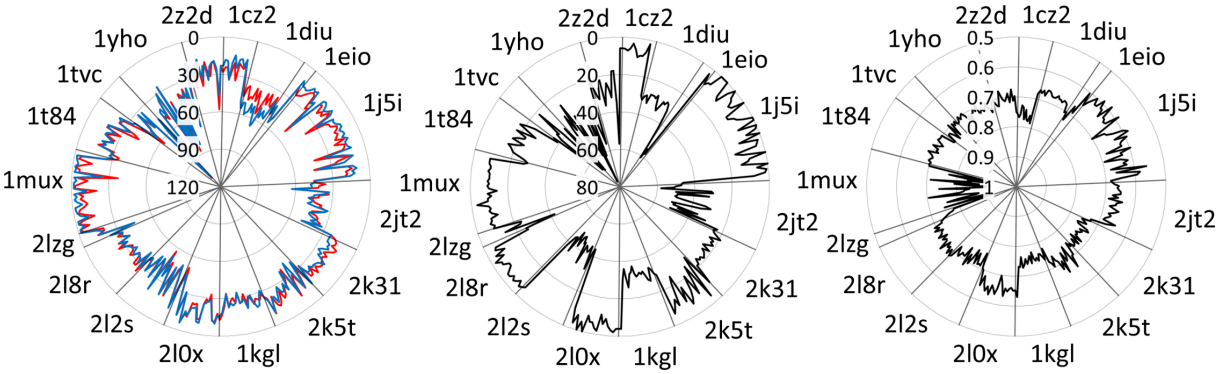


Figure 4.2.5. Binding site properties of structural ensembles used to generate dataset 2. Different binding site descriptors are depicted in the radar plots together with the corresponding PDB-IDs of the NMR ensembles. Left: The number of binding site acceptor (red) and donor (blue) atoms in the ligand binding sites of different NMR ensembles. Centre: The number of aromatic atoms within the models' binding sites. Right: The hydrophobicity of different NMR model binding sites.

The performance differences between Cavbase and RAPMAD are not that pronounced for the NMR structures when compared to the results for the structures with identical sequences. The poor performance of both tools relates to the fact that binding sites are extracted based on a LIGSITE[273] detection. Therefore, not all cavities of interest are necessarily identified and considered for the comparison. This represents a major drawback of both methods and is reflected in the progression of the ROC curve (linear slope in the terminal part of the curve). Nonetheless, both methods show a convincing early enrichment and might be beneficial for identifying highly similar binding sites in a drug repurposing or polypharmacology project. The exclusion of all pairs that could not be compared by the aforementioned tools for benchmark analyses leads to improved performance. However, this improvement is negligible, and the problem will persist for other structures.

It might be beneficial to use MD-derived structures for a screening of a large binding site library with the binding site of interest for the elucidation of potential off-targets when applying these tools. Especially, off-target predictions and the analysis of evolutionary relationships should not solely rely on Cavbase, RAPMAD, or VolSite/Shaper.

The same holds true for Grim, TIFP, and IsoMIF. Remarkably, all these tools are interaction-based methods leading to the question of whether unstable interactions of the ligand with the POI vanish in some ensemble structures. The underlying NMR structures were analysed and the comparison of three structures from the NMR ensemble with the PDB-ID 1eio illustrates that only a small fraction of all observed interactions remains stable in solution (Figure 4.2.6). The figure also illustrates the changes in the binding site environment. These changes do not influence residue- and surface-based methods to a major extent, but they hamper a comparison that relies on distinct sets of interactions. This finding also indicates that the inclusion of interaction strengths (e.g., as weights) for an interaction-based comparison method might prove beneficial and could lead to improved performance. Again, the inclusion of various protein-ligand complexes for the cavity comparison with an interaction-based tool is inevitable for the success of the study. Especially the outcome of IsoMIF is considerably affected by the binding site flexibility. Therefore, the tool should be used in combination with various structures to guarantee reliable results. Grim and TIFP show similar performance with respect to the AUC. Nevertheless, the graph-based method outperforms TIFP in terms of early enrichment.

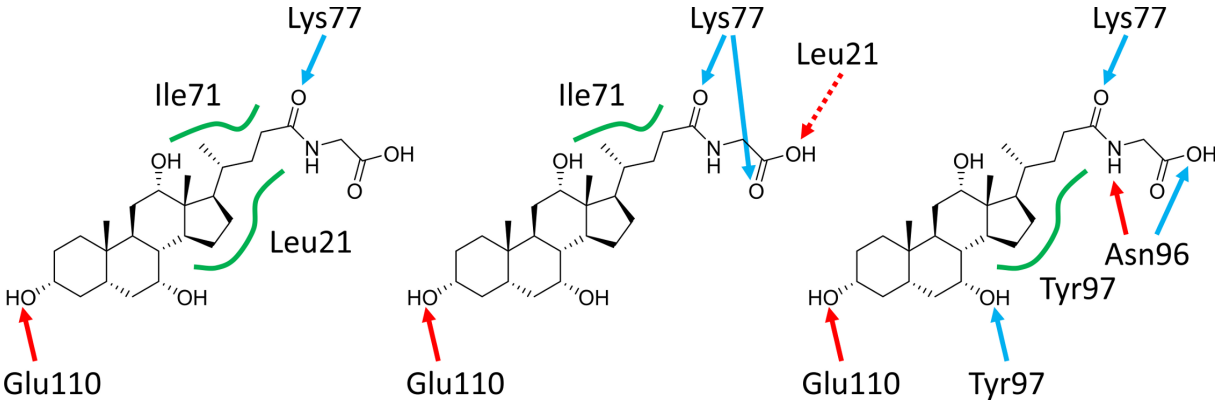
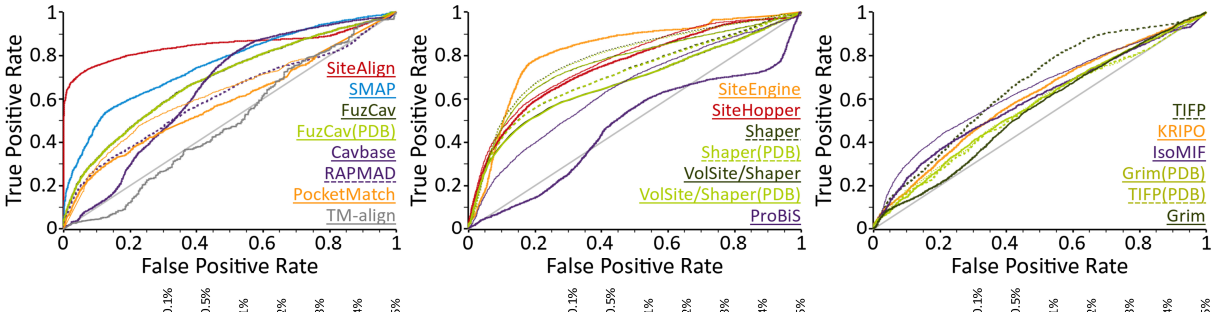


Figure 4.2.6. Different interaction patterns for structures from the solution NMR ensemble of ileal lipid-binding protein. The ensemble (PDB-ID 1eio, chain A) contains five conformers. Models 3, 4, and 5 (from left to right) were used to generate this illustration with the help of MOE[257] and PoseView.[385] Residues with alternating interaction patterns (red arrows: hydrogen bond acceptors, blue arrows: hydrogen bond donors, green outlines: hydrophobic interaction partners) in the different conformations of the binding site occupy nearly half of the pocket. The remaining part of the pocket is mainly engaged in hydrophobic contacts with the small molecule glycocholate.

In summary, we can pinpoint some methods (SMAP, TM-align, SiteAlign, SiteEngine, and ProBiS) which do not depend on the binding site conformation. These tools should be preferentially used for query-based investigations or the calculation of similarity matrices for a restricted set of protein structures. Otherwise, only the generation of structural ensembles of the POI will lead to meaningful comparison outcomes for methods that are sensitive to changes due to the binding site dynamics.

Decoy Structures

The recovery of binding site pairs of structures with identical sequences or NMR models highlighted some basic differences in the sensitivity of the methods with respect to binding site definition and flexibility. However, the question arises whether the analysed algorithms can distinguish between structures with several substitutions and the original structures. This puts up a major challenge with respect to the scoring schemes of the methods. The dataset of structures with identical sequences built the foundation for this evaluation. For the separate groups, one structure each was used to generate binding site variants with one, two, three, four, or five substitutions by residues which differ in their physicochemical properties (dataset 3) or their size (dataset 4). The first dataset of decoy binding sites (dataset 3) was developed to assess the ability of the tools to distinguish binding sites with different physicochemical properties. For this dataset, most tools failed to distinguish between the original binding site and variants with one to five substitutions (see Figure A2.6 in Appendix 2 for the AUC values and EFs for all methods). Therefore, it was analysed whether the structures with identical sequences were ranked higher than these with five residue substitutions. The respective ROC curves and EFs are depicted in Figure 4.2.7.



	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	1.4	1.7	1.4	1.1	1.1	1.0	1.0
FuzCav	1.9	2.0	2.0	1.9	1.9	1.8	1.8
FuzCav (PDB)	1.9	2.0	2.0	1.9	1.9	1.8	1.8
PocketMatch	0.0	0.0	0.9	1.4	1.6	1.6	1.6
RAPMAD	1.9	2.0	2.0	1.7	1.7	1.7	1.7
SiteAlign	1.9	2.0	2.0	2.0	2.0	2.0	2.0
SMAP	1.2	1.8	1.9	1.9	1.9	2.0	1.9
TM-align	0.0	0.0	0.0	0.8	0.9	0.8	0.8

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
ProBiS	1.7	1.5	1.2	1.0	1.0	1.0	1.0
Shaper	1.9	2.0	2.0	1.6	1.6	1.6	1.7
Shaper (PDB)	1.9	2.0	2.0	1.6	1.6	1.6	1.6
SiteEngine	1.9	2.0	2.0	1.8	1.6	1.5	1.5
SiteHopper	1.9	2.0	2.0	1.9	1.8	1.8	1.8
VolSite/Shaper	1.9	2.0	2.0	1.7	1.7	1.7	1.7
VolSite/Shaper (PDB)	1.9	2.0	2.0	1.7	1.7	1.7	1.7

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Grim	1.6	1.3	1.2	1.3	1.4	1.4	1.4
Grim (PDB)	1.9	2.0	1.8	1.6	1.5	1.4	1.3
IsoMIF	1.9	2.0	2.0	1.6	1.5	1.5	1.5
KRIPO	1.9	2.0	2.0	1.6	1.6	1.5	1.5
TIFP	1.7	2.0	2.0	1.6	1.5	1.5	1.5
TIFP (PDB)	0.0	1.1	1.5	1.6	1.5	1.4	1.3

Figure 4.2.7. Evaluation of different binding site comparison tools with respect to dataset 3 (five substitutions). Top: ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is coloured according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. PocketMatch showed the best AUC for the score $PMScore_{min}$ (thin orange line). The scores SVA, RefTversky (color), RefTversky (color), RefTversky (color), RefTversky (color), and ColorTanimoto led to the highest AUC values for ProBiS, Shaper, Shaper(PDB), VolSite/Shaper, VolSite/Shaper(PDB), and SiteHopper, respectively (thin lines). The highest AUC was obtained for IsoMIF and TIFP(PDB) if taniM and the Tanimoto coefficient were used as the similarity measures (thin lines). Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

Intriguingly, the methods SiteAlign, SMAP, SiteEngine, SiteHopper, and Shaper were significantly superior to all interaction-based methods in terms of the AUC. Nevertheless, the early enrichment of the active pairs is comparably high. The results for TM-align are explicable when considering that the TM-score does not consider the residue type.

Cavbase, Grim, ProBiS, and TIFP show the lowest early enrichment. When taking one, two, three, four, and five residue substitutions into account this trend becomes even more pronounced (see Figure A2.6 in Appendix 2). Obviously, these tools assign high similarity scores to the matches between the original structure and its decoy structures, whereas other structures with different binding site conformations are regarded as being more dissimilar.

To further investigate the similarity scoring performance, it was evaluated whether the methods are able to rank the matches between the original structure and its variants with respect to the number of substitutions. To this end, the Spearman's Rho rank correlation coefficients between the number of substitutions and the corresponding similarity score to the original structure were calculated. The outcome of this evaluation is shown in Table 4.2.3. The results for TM-align were omitted as the method is not able to distinguish between the original structure and its variants due to the missing incorporation of residue properties.

Negative correlation coefficients indicate a decreasing similarity score with an increasing number of mutations which is to be expected for binding site comparison tools which can discriminate between binding sites with different physicochemical properties. The value -1 indicates a perfect correlation. The outcome of this analysis shows that SMAP, SiteHopper, and KRIPO as residue-, surface-, and interaction-based methods are best suited to distinguish between similar and dissimilar binding sites. SiteAlign and SiteEngine provided nearly equally good results. For Cavbase and RAPMAD, no results were obtained for the structure with the PDB-ID 4ca7 (chain A) as the implemented LIGSITE[273] algorithm for the binding site assignment did not detect this buried binding site (Figure A2.7 in Appendix 2).

The results obtained with IsoMIF for some structures show a very weak correlation between the number of mutations and the similarity score. The results are comparable to those of TIFP and Grim which were outperformed by IsoMIF for the two previously discussed datasets. The same holds true for PocketMatch and ProBiS. In some cases, no correlation at all is discernible.

The second dataset of decoy binding sites was compiled to evaluate the influence of the cavity shape on the outcome of this study. Randomly chosen residues were replaced by residues with a different shape. The benchmark results for this dataset are depicted in Figure 4.2.8. A different trend in comparison to dataset 3 is that most surface-based methods are characterised by a better performance. This holds especially true for Shaper and VolSite/Shaper. The introduction of differently sized residues causes these improvements. In contrast, the performance of ProBiS and the interaction-based methods did not improve. Strikingly, the performance of SMAP is worse than for dataset 3. This tool seems to be more sensitive towards the chemical properties of the residues than towards their shape. A plausible explanation is that SMAP models geometric binding site features based on the Cα coordinates. These do not change upon the introduction of the substitutions. A calculation of the Pearson correlation between score and number of substitutions (Table 4.2.4) shows that the results of PocketMatch, ProBiS, Grim, IsoMIF, and TIFP did not improve upon the introduction of differently sized residues. This is in line with the previous finding that their score ranges for similar and dissimilar sites partially overlap (see Figure A2.5 in Appendix 2).

Table 4.2.3. The similarity ranking of physicochemically different sites. Spearman's Rho correlation coefficients for a comparison between crystal structures and their corresponding decoy structures with one, two, three, four, and five substitutions of randomly chosen residues for residue-based, surface-based, and interaction-based binding site comparison methods are given. A colour gradient was applied to highlight a prominent negative correlation (green), a correlation coefficient of -0.5 (yellow), no correlation or a positive correlation (red).

Residue-Based Methods

PDB-ID. Chain	*Cavbase*	*FuzCav*	*FuzCav (PDB)*	*PocketMatch*	*RAPMAD*	*SiteAlign*	*SMAP*
1kmv.A	-0.90	-0.90	-0.90	-0.31	-0.58	-0.95	-0.89
1odm.A	-0.79	-0.90	-0.90	-0.34	-0.71	-0.89	-0.91
2qxw.A	-0.67	-0.79	-0.79	-0.53	-0.61	-0.75	-0.81
3f17.A	-0.67	-0.87	-0.87	-0.59	-0.69	-0.88	-0.93
3rm2.H	-0.53	-0.58	-0.58	-0.52	-0.60	-0.81	-0.80
3t10.A	-0.72	-0.56	-0.57	-0.52	-0.65	-0.75	-0.92
3u5l.A	-0.66	-0.83	-0.83	-0.59	-0.78	-0.88	-0.87
3u9w.A	-0.86	-0.72	-0.72	-0.33	-0.71	-0.90	-0.82
4bfz.A	-0.70	-0.68	-0.68	-0.31	-0.54	-0.71	-0.84
4buu.A	-0.63	-0.72	-0.72	-0.56	-0.75	-0.76	-0.92
4ca7.A	-	-0.69	-0.69	-0.48	-	-0.84	-0.96
4fpt.A	-0.65	-0.86	-0.86	-0.57	-0.61	-0.85	-0.94

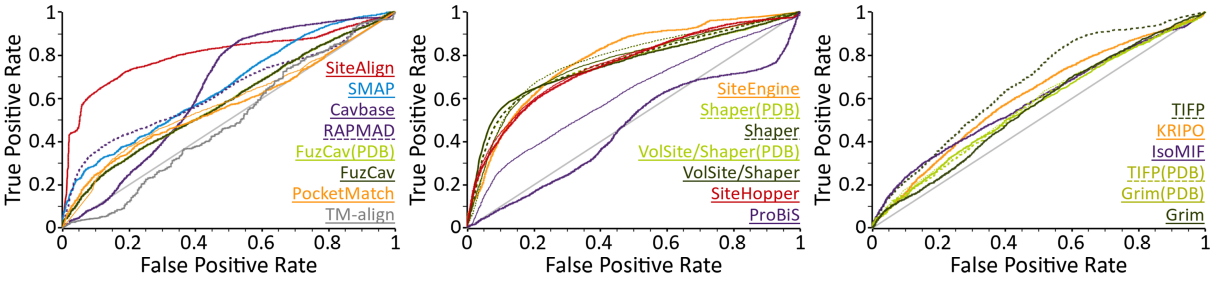
Surface-Based Methods

PDB-ID. Chain	*ProBiS*	*Shaper*	*Shaper (PDB)*	*SiteEngine*	*SiteHopper*	*VolSite/ Shaper*	*VolSite/ Shaper (PDB)*
1kmv.A	-0.53	-0.89	-0.89	-0.76	-0.90	-0.91	-0.91
1odm.A	-0.01	-0.59	-0.59	-0.71	-0.95	-0.92	-0.92
2qxw.A	-0.26	-0.71	-0.70	-0.76	-0.85	-0.73	-0.69
3f17.A	-0.70	-0.69	-0.69	-0.71	-0.77	-0.70	-0.71
3rm2.H	-0.52	-0.60	-0.60	-0.62	-0.79	-0.50	-0.50
3t10.A	-0.31	-0.65	-0.64	-0.66	-0.82	-0.63	-0.62
3u5l.A	-0.54	-0.67	-0.67	-0.65	-0.94	-0.72	-0.73
3u9w.A	0.02	-0.72	-0.72	-0.81	-0.84	-0.67	-0.67
4bfz.A	-0.38	-0.73	-0.74	-0.81	-0.90	-0.77	-0.77
4buu.A	-0.40	-0.73	-0.71	-0.65	-0.90	-0.65	-0.69
4ca7.A	0.22	-0.31	-0.32	-0.75	-0.87	-0.63	-0.63
4fpt.A	0.43	-0.83	-0.82	-0.85	-0.93	-0.85	-0.85

Interaction-Based Methods

PDB-ID. Chain	*Grim*	*Grim (PDB)*	*IsoMIF*	*KRIPO*	*TIFP*	*TIFP (PDB)*
1kmv.A	-0.70	-0.72	-0.56	-0.67	-0.35	-0.70
1odm.A	-0.66	-0.57	-0.72	-0.77	-0.64	-0.61
2qxw.A	-0.73	-0.62	-0.60	-0.75	-0.62	-0.57
3f17.A	-0.58	-0.53	-0.52	-0.75	-0.59	-0.28
3rm2.H	-0.61	-0.37	-0.68	-0.72	-0.42	-0.28
3t10.A	-0.46	-0.42	-0.56	-0.61	-0.30	-0.43
3u5l.A	-0.60	-0.56	-0.24	-0.68	-0.45	-0.59
3u9w.A	-0.68	-0.73	-0.65	-0.73	-0.50	-0.54
4bfz.A	-0.51	-0.50	-0.61	-0.51	-0.45	-0.47
4buu.A	-0.57	-0.32	-0.20	-0.65	-0.37	-0.32
4ca7.A	-0.67	-0.43	-0.79	-0.63	-0.44	-0.41
4fpt.A	-0.72	-0.76	-0.40	-0.80	-0.54	-0.61

The superiority of SiteAlign compared to all other tools is pronounced for both datasets, 3 and 4. In consequence, a potential applicability domain for SiteAlign is the accurate evaluation of minor binding site differences. This is of interest in polypharmacology studies and drug repurposing projects, as well as for the evaluation of off-target effects. The same holds true for SMAP and SiteHopper although they show a lower early enrichment. These tools do not necessarily require structural ensembles as input as their sensitivity with respect to the binding site flexibility is low. Therefore, no additional computational cost arises for the application of these methods. Approaches which provide such accurate scoring measures might be suitable choices to rescore potential matches of similar binding sites obtained with the help of a faster, but less accurate binding site comparison tool.



	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	1.7	1.6	1.6	1.1	1.2	1.0	1.0
FuzCav	1.5	1.9	2.0	1.8	1.6	1.5	1.5
FuzCav (PDB)	1.5	1.9	2.0	1.8	1.6	1.5	1.5
PocketMatch	0.0	0.0	0.8	1.4	1.6	1.6	1.6
RAPMAD	1.9	2.0	2.0	1.7	1.7	1.7	1.7
SiteAlign	1.9	2.0	2.0	2.0	1.9	1.9	1.9
SMAP	1.2	1.8	1.9	1.9	1.9	1.9	1.8
TM-align	0.0	0.0	0.0	0.8	0.9	0.8	0.8

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
ProBiS	1.6	1.7	1.3	0.9	0.9	1.0	0.9
Shaper	1.9	2.0	2.0	1.8	1.7	1.7	1.7
Shaper (PDB)	1.9	2.0	2.0	1.7	1.7	1.7	1.7
SiteEngine	1.9	2.0	2.0	1.8	1.8	1.8	1.8
SiteHopper	1.9	2.0	2.0	1.8	1.8	1.8	1.8
VolSite/Shaper	1.9	2.0	2.0	2.0	1.9	1.9	1.8
VolSite/Shaper (PDB)	1.9	2.0	2.0	1.9	1.9	1.9	1.8

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Grim	1.6	1.3	1.4	1.4	1.4	1.4	1.4
Grim (PDB)	1.9	1.8	1.8	1.6	1.6	1.4	1.3
IsoMIF	1.9	2.0	2.0	1.7	1.6	1.6	1.6
KRIPO	1.9	2.0	2.0	1.6	1.4	1.4	1.3
TIFP	0.8	1.8	1.9	1.7	1.6	1.6	1.5
TIFP (PDB)	0.0	1.0	1.5	1.6	1.5	1.4	1.4

Figure 4.2.8. Evaluation of different binding site comparison tools with respect to dataset 4 (five substitutions). Top: ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is coloured according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. PocketMatch showed the best AUC for the score $PMScore_{min}$ (thin orange line). The scores SVA, RefTversky (color), RefTversky (color), RefTversky (color), RefTversky (color), and ColorTanimoto led to the highest AUC values for ProBiS, Shaper, Shaper(PDB), VolSite/Shaper, VolSite/Shaper(PDB), and SiteHopper, respectively (thin lines). The highest AUC was obtained for TIFP(PDB) with the Tanimoto coefficient as the similarity metric. Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

A closer examination of the Spearman's Rho correlation coefficients (Table 4.2.4) shows that the SMAP and SiteHopper scoring schemes are best suited to estimate the degree of shape dissimilarity by the similarity score. For the interaction-based methods, KRIPO led to the clearest correlation between the number of substitutions and the score. Consequently, the application of these tools might offer an opportunity to correlate ligand affinities with binding site similarities, e.g., for protein kinases as previously shown with other methods.[386–388] A clear differentiation between minor dissimilarities is an essential necessity for such analyses and the aforementioned tools might be helpful to reliably rank binding sites according to their similarities. However, if a robust method is needed that should not be influenced by minor dissimilarities, one of the worse performing tools might be a better choice.

Table 4.2.4. The similarity ranking of sites with different shapes. Spearman's Rho correlation coefficients for a comparison between crystal structures and their corresponding decoy structures with one, two, three, four, and five substitutions of randomly chosen residues for residue-based, surface-based, and interaction-based binding site comparison methods are given. A colour gradient was applied to highlight a prominent negative correlation (green), a correlation coefficient of -0.5 (yellow), no correlation or a positive correlation (red).

Residue-Based Methods

PDB-ID. Chain	*Cavbase*	*FuzCav*	*FuzCav (PDB)*	*PocketMatch*	*RAPMAD*	*SiteAlign*	*SMAP*
1kmv.A	-0.89	-0.67	-0.67	-0.55	-0.52	-0.83	-0.90
1odm.A	-0.94	-0.73	-0.73	-0.31	-0.66	-0.54	-0.92
2qxw.A	-0.68	-0.69	-0.69	-0.19	-0.57	-0.74	-0.93
3f17.A	-0.40	-0.55	-0.55	-0.48	-0.55	-0.82	-0.89
3rm2.H	-0.60	-0.60	-0.60	-0.66	-0.61	-0.73	-0.90
3t10.A	-0.77	-0.44	-0.44	-0.57	-0.61	-0.59	-0.92
3u5l.A	-0.53	-0.62	-0.62	-0.65	-0.32	-0.64	-0.92
3u9w.A	-0.87	-0.79	-0.79	-0.40	-0.52	-0.77	-0.88
4bfz.A	-0.89	-0.28	-0.28	-0.52	-0.77	-0.82	-0.89
4buu.A	-0.72	-0.63	-0.63	-0.38	-0.50	-0.52	-0.92
4ca7.A	-	-0.57	-0.57	-0.49	-	-0.81	-0.90
4fpt.A	-0.79	-0.61	-0.61	-0.51	-0.36	-0.51	-0.88

Surface-Based Methods

PDB-ID. Chain	*ProBiS*	*Shaper*	*Shaper (PDB)*	*SiteEngine*	*SiteHopper*	*VolSite/ Shaper*	*VolSite/ Shaper (PDB)*
1kmv.A	-0.74	-0.90	-0.90	-0.85	-0.92	-0.90	-0.90
1odm.A	0.08	-0.86	-0.86	-0.88	-0.95	-0.89	-0.90
2qxw.A	-0.54	-0.73	-0.74	-0.86	-0.87	-0.70	-0.76
3f17.A	-0.18	-0.66	-0.65	-0.79	-0.74	-0.65	-0.64
3rm2.H	-0.34	-0.47	-0.46	-0.60	-0.81	-0.40	-0.40
3t10.A	-0.47	-0.66	-0.66	-0.74	-0.85	-0.49	-0.49
3u5l.A	-0.42	-0.75	-0.79	-0.76	-0.94	-0.84	-0.84
3u9w.A	0.28	-0.73	-0.72	-0.78	-0.81	-0.57	-0.57
4bfz.A	-0.26	-0.70	-0.71	-0.84	-0.93	-0.72	-0.68
4buu.A	-0.58	-0.59	-0.62	-0.85	-0.83	-0.48	-0.47
4ca7.A	0.27	-0.43	-0.44	-0.76	-0.86	-0.46	-0.47
4fpt.A	0.07	-0.80	-0.80	-0.89	-0.86	-0.84	-0.84

Interaction-Based Methods

PDB-ID. Chain	*Grim*	*Grim (PDB)*	*IsoMIF*	*KRIPO*	*TIFP*	*TIFP (PDB)*
1kmv.A	-0.64	-0.64	-0.69	-0.72	-0.50	-0.63
1odm.A	-0.57	-0.68	-0.79	-0.76	-0.54	-0.65
2qxw.A	-0.75	-0.66	-0.66	-0.61	-0.54	-0.66
3f17.A	-0.58	-0.40	-0.54	-0.52	-0.49	-0.37
3rm2.H	-0.61	-0.43	-0.61	-0.72	-0.51	-0.41
3t10.A	-0.37	-0.47	-0.62	-0.70	-0.35	-0.32
3u5l.A	-0.71	-0.48	-0.29	-0.68	-0.50	-0.33
3u9w.A	-0.70	-0.61	-0.75	-0.74	-0.55	-0.60
4bfz.A	-0.57	-0.40	-0.77	-0.79	-0.32	-0.33
4buu.A	-0.62	-0.63	-0.17	-0.74	-0.52	-0.49
4ca7.A	-0.47	-0.44	-0.51	-0.68	-0.36	-0.43
4fpt.A	-0.74	-0.68	-0.56	-0.72	-0.70	-0.75

A potential application for tools that are insensitive to the introduced substitutions is the elucidation of non-obvious binding site similarities, as shown for SiteAlign,[370] ProBiS,[367] and KRIPO.[361] Examples for such unapparent similarities are given in the following sections.

Taken together, the results indicate that several residue-based (but especially SiteAlign) and surface-based tools, as well as the interaction-based method KRIPO, are suited to accurately score similarities between binding site pairs which differ due to only small modifications. Depending on the desired outcome (identification of minor dissimilarities or the identification of only moderately related sites), the most suitable method should be chosen with care.

The ROCS Dataset

Similar Ligands – Similar Binding Sites?

In addition to these fundamental datasets, a set of sequentially and structurally unrelated proteins that bind to similar ligands was prepared. As it was shown before, the binding modes of ligands with a high 2D similarity can considerably vary depending on the protein binding site. This is partially due to a high degree of flexibility of the ligand of interest. The OpenEye tool ROCS[264] was used to identify protein-bound ligands that do not only share common physicochemical properties but also a similar shape. Based on the ligand files as stored in the sc-PDB[10] a ROCS-based all-against-all comparison was performed. Highly distinct proteins that nevertheless bind to very similar ("active" pairs) or highly dissimilar ("inactive" pairs) ligands with respect to the ComboTanimoto were extracted. The resulting protein pairs were analysed using TM-align to exclude pairs with a high sequence and structure similarity. All pairs with a TM-score above 0.3 were excluded from the dataset, while the remaining pairs were used as input structures for the binding site comparison tools. The final dataset includes 15,339 active pairs and 56,179 inactive pairs for benchmarking purposes.

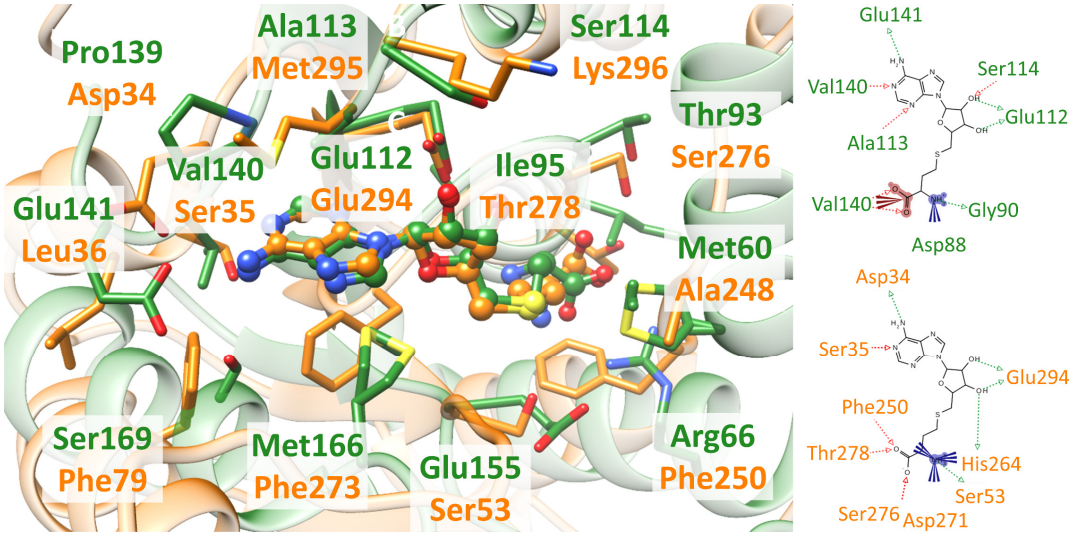
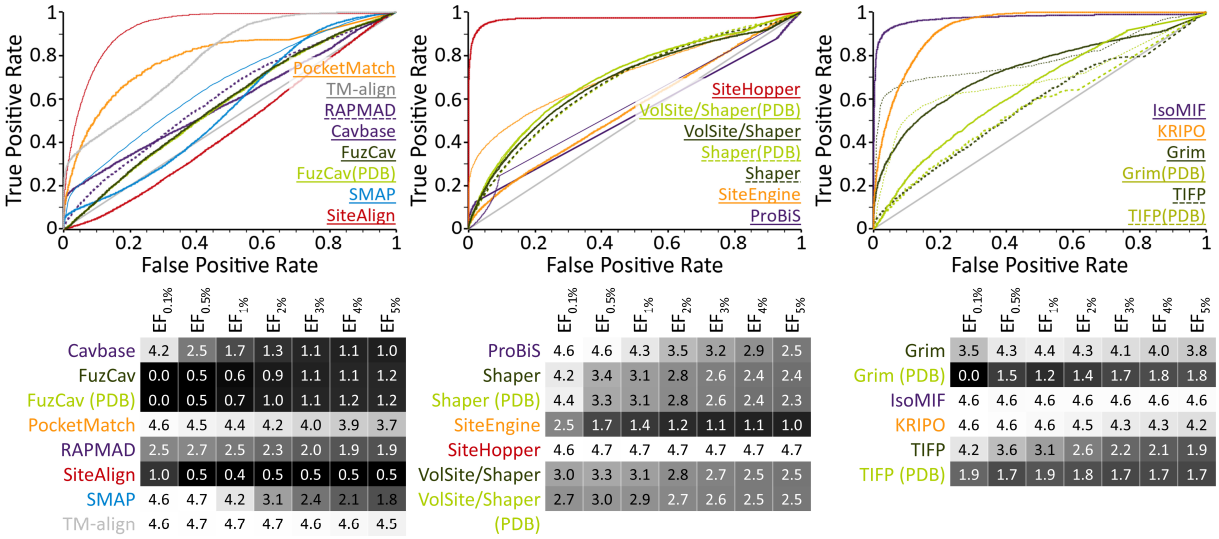


Figure 4.2.9. A non-obvious binding site similarity. Left: Binding site alignment of the human enzyme arginine N-methyltransferase 6 (PDB-ID 4hc4, chain A, green) and cytosine-N4-specific methyltransferase from *Proteus vulgaris* (PDB-ID 1boo, chain A, orange) with SAH in ball-and-stick representation. The corresponding binding site residues are coloured according to the structure. The figure was generated with UCSF Chimera.[228] Right: The interaction patterns of SAH in the binding sites of the complexes with the PDB-IDs 4hc4 (top) and 1boo (bottom) are shown to highlight crucial interaction-based similarities between both sites. The figures were generated with LigandScout 4.0.[383]

High similarities in the shape and physicochemical properties of the underlying ligands might indicate similar interaction patterns in the binding sites of interest. Therefore, this dataset should disclose the benefits of interaction-based binding site comparison tools. This is presented for an example in Figure 4.2.9. Both protein cavities bind to *S*-adenosyl-*L*-homocysteine (SAH) in a similar conformation and both ligands share common interaction patterns. Nonetheless, the binding site environment is different. In consequence, residue- and surface-based comparison methods will probably not assign high similarity scores for this binding site pair, although it is a promising example of related sites. This basic assumption is verified by the outcome of the analysis as presented in Figure 4.2.10.



	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	4.2	2.5	1.7	1.3	1.1	1.1	1.0
FuzCav	0.0	0.5	0.6	0.9	1.1	1.1	1.2
FuzCav (PDB)	0.0	0.5	0.7	1.0	1.1	1.2	1.2
PocketMatch	4.6	4.5	4.4	4.2	4.0	3.9	3.7
RAPMAD	2.5	2.7	2.5	2.3	2.0	1.9	1.9
SiteAlign	1.0	0.5	0.4	0.5	0.5	0.5	0.5
SMAP	4.6	4.7	4.2	3.1	2.4	2.1	1.8
TM-align	4.6	4.7	4.7	4.7	4.6	4.6	4.5

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
ProBiS	4.6	4.6	4.3	3.5	3.2	2.9	2.5
Shaper	4.2	3.4	3.1	2.8	2.6	2.4	2.4
Shaper (PDB)	4.4	3.3	3.1	2.8	2.6	2.4	2.3
SiteEngine	2.5	1.7	1.4	1.2	1.1	1.1	1.0
SiteHopper	4.6	4.7	4.7	4.7	4.7	4.7	4.7
VolSite/Shaper	3.0	3.3	3.1	2.8	2.7	2.5	2.5
VolSite/Shaper (PDB)	2.7	3.0	2.9	2.7	2.6	2.5	2.5

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Grim	3.5	4.3	4.4	4.3	4.1	4.0	3.8
Grim (PDB)	0.0	1.5	1.2	1.4	1.7	1.8	1.8
IsoMIF	4.6	4.6	4.6	4.6	4.6	4.6	4.6
KRIPO	4.6	4.6	4.6	4.5	4.3	4.3	4.2
TIFP	4.2	3.6	3.1	2.6	2.2	2.1	1.9
TIFP (PDB)	1.9	1.7	1.9	1.8	1.7	1.7	1.7

Figure 4.2.10. Evaluation of different binding site comparison tools with respect to dataset 5. Top: ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is coloured according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. The scoring measures that yielded the highest AUC were distance d1 and the Tanimoto coefficient for SiteAlign (thin red line) and SMAP (thin blue line), respectively. VolSite/Shaper, VolSite/ Shaper(PDB), and SiteEngine showed the highest AUC values when using the similarity scores Tanimoto (fit), Tanimoto (fit), and DistanceScore, respectively (thin lines). TIFP AUC values significantly improved for the Hamming distance as the distance measure (thin dark green line). Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

Apart from SiteHopper, the interaction-based method IsoMIF clearly outperforms the remaining binding site comparison tools. Interestingly, a switch from a random scoring of similar and dissimilar binding site pairs to a reasonable differentiation between both types was noticed if the scoring scheme of SiteAlign is changed. The scoring scheme which proved successful for this set of cavities uses a distance measure that detects global binding site similarities (d1). In contrast, the distance measure that was most appropriate for the preceding dataset (d3) allows for the identification of local binding site similarities. Additionally, it uses the distance to the centre of the side-chain (Ccentroid) rather than that to the rotamer's Cβ atom. This finding underlines a peculiarity of this dataset. In comparison to all previously applied datasets, the ligands of the "active" site pairs overlap in shape and physicochemical properties, and their size is comparable.

Therefore, the binding site definition allows for similarly excised binding site sections whose global similarity will be high. A plausible conclusion is that the use of distance measure d1 in SiteAlign is best suited for the detection of similarities between protein structures in complex with similar ligands. The use of the distance measure d3 would hamper their identification as it will be ranked lower than other site pairs with a low global, but a high local similarity.

The inferior performance of FuzCav is also easily explicable when having a look at Figure 4.2.9. The side-chains of the hydrophobic residues Met166 and Phe273 are located in a similar position with respect to the ligand. Nevertheless, their backbone atoms are part of different backbone structures in opposite locations of the binding site. In contrast to the previously investigated datasets, this is a unique feature of this set of similar site pairs and prevents the success of residue-based methods that make use of Cα atom positions for the assignment of binding site features. This can also be demonstrated by the poor performance of SMAP which gave good to excellent results for the other datasets.

A more striking observation for this dataset is the obviously high early enrichment observed for TM-align. Partially, this can be explained by the rigorous binding site definition based on the distance to the ligands whose shape is highly similar to their “active” counterpart. On the other hand, the algorithm also relies on Cα atom coordinates and an initial secondary structure-based alignment. It can also be excluded that the underlying proteins show a high overall similarity as TM-align was used to eliminate these protein pairs. The analysis of the pair with the highest TM-score in the dataset revealed the explanation for this observation. Although both proteins differ in their overall structure if all chains are considered, both structures have a similar protein chain belonging to the same CATH homology superfamily (P-loop containing nucleotide triphosphate hydrolase 3.40.50.300 for the structures with the PDB-IDs 4hdq (chain B) and 3tnf (chain A)). Two protein chains of the structure with the PDB-ID 3tnf are involved in ligand binding and were therefore retained for the structure and binding site comparison. Nevertheless, such protein pairs do not influence the interpretation of the results as both binding sites differ with respect to the underlying residues. To answer the question of whether similar issues could be observed for the decoy structures, the inactive pairs with high TM-scores were analysed. No such case was found within the dataset pairs. In regard to this finding, it is even more surprising that most residue-based methods were not able to correctly distinguish similar and dissimilar pairs. The exceptions are SiteAlign in combination with distance score d1 and PocketMatch. The same conclusion holds true for the surface-based methods except for SiteHopper which clearly outperforms all residue- and surface-based methods. The latter finding can be attributed to the algorithmic similarities between ROCS, which was used to identify the ligand-based similarities, and SiteHopper. They share a common approach to shape matching which biases the results for this dataset.

Most interestingly, the IsoMIF-derived similarity scores lead to an AUC of nearly 1, the early enrichment of IsoMIF and KRIPO is comparable. Of all interaction-based tools, these approaches seem to be most appropriate to correctly identify close relationships between binding sites occupied with similar ligands.

An Inevitable Bias?

The outcome of the analysis was unanticipated as all analysed tools showed satisfactory performance on different datasets of interesting binding site similarities (see also below for dataset 8). The clear superiority of IsoMIF and SiteHopper was not seen before. SiteHopper showed good scoring performance, but its approach is highly related to the ROCS approach used to generate the dataset. To investigate whether the ligand-based similarities indeed correspond to similarities of the respective binding sites, the similarity scores for the binding site pairs obtained with SiteHopper were considered. In the earlier analyses, the performance of SiteHopper was not only convincing with respect to binding site definition and flexibility, but also showed a good scoring performance. For most pairs, the shape similarity was much higher than the Tanimoto coefficient for common physicochemical properties which is in accordance with the analyses of the ligand-based similarities. In general, only one-third of the binding site similarities result in a ColorTanimoto above 0.15.

An analysis of the binding site descriptors (using fpocket[8]) of all structures that are exclusively characterised by high TanimotoShape-based and TanimotoColor-based (i.e., physicochemical) similarity and those that show a high physicochemical similarity reveals a distinct trend. Binding sites which bind to similar ligands and share mainly common shape properties are characterised by a high fpocket hydrophobicity and low fpocket polarity score as compared to those that bind to similar ligands and show a high physicochemical similarity. This suggests that the overall superiority of IsoMIF and SiteHopper results from scores derived from a high shape, but not physicochemical similarity.

This is exemplified by some interesting similarities between ligands of unrelated enzymes based on the ROCS comparison. Fludrocortisone, whose structure is known in complex with a mutant human androgen receptor (PDB-ID 1gs4), shows a high shape and colour similarity towards dexamethasone-21-phosphate bound to the Pin1 substrate binding domain (PDB-ID 3tca). Nevertheless, the K_d of 4.4 mM illustrates the low affinity of the ligand towards Pin1 which mainly interacts by its phosphate moiety. Another similarity was found between sulfathiazole in complex with sepiapterin reductase (PDB-ID 4j7u) and 2,4,5-trichloro-*N*-(5-methyl-1,2-oxazol-3-yl)benzenesulfonamide as a weak thumb pocket 2 inhibitor of the HCV NS5B polymerase (PDB-ID 4iz0, IC_{50} = 100 µM). This hit resulted from a fragment-based screening approach. The same holds true for 3-methyl-3,4-dihydroquinazolin-2(1*H*)-one in complex with the N-terminal bromodomain of Brd2 (PDB-ID 4a9e, 12% inhibition at a concentration of 100 µM[389]) whose binding mode shows a high shape and physicochemical similarity to that of chlorzoxazone bound to nitric oxide synthase (PDB-ID 1m8d, K_d = 4.3 µM). Moreover, a similarity between a peptide-mimetic ligand bound to *E. coli* peptide deformylase (PDB-ID 3k6l) and marimastat in complex with ADAMTS-1 (PDB-ID 2jih) was identified. Most of these binding site pairs are characterised by a comparatively low binding site similarity. Their SiteHopper scores result from a remarkably high shape, but low physicochemical similarity (1gs4 and 3tca: ShapeTanimoto = 0.24 and ColorTanimoto = 0.10, 4j7u and 4iz0: ShapeTanimoto = 0.35 and ColorTanimoto = 0.11, 4a9e and 1m8d: ShapeTanimoto = 0.36 and ColorTanimoto = 0.11, 3k6l and 2jih: ShapeTanimoto = 0.31 and ColorTanimoto = 0.2).

For binding sites complexed with flavone scaffolds that show high ligand-based similarity, the overall similarity between the binding sites is also low. Most similarities fall below the threshold for significant binding site similarities. For sites binding to related flavone-scaffolds, distinct interaction patterns are found that convey selectivity. This is shown for the similarity between tankyrase 2 (PDB-ID 4hkn) and dihydroflavonol-4-reductase (PDB-ID 2nnl) in Figure 4.2.11 (left panel). However, the ROC curve for SiteHopper[272] given in Figure 4.2.10 suggests that binding site similarity can lead to the identification of binding site pairs which mainly share common shape properties. The comparison of both sites with IsoMIF led to a score of 0.7 which is indicative of highly similar sites. Therefore, the analysis of the similarity scores in combination with the binding site descriptors is inevitable for further investigations and the validation of detected similarities. A visualisation of the binding site alignment and the corresponding ligand-based alignment underlines the differences in the interaction of both molecules towards their respective targets.

In contrast, the similarity between *E. coli* peptide deformylase and ADAMTS-1 (both metalloproteases) is higher (ShapeTanimoto = 0.31, ColorTanimoto = 0.2). An examination of all similarities to ADAMTS-1 (PDB-ID 2jih) in complex with marimastat shows pronounced similarities to peptide deformylase enzymes of different pathogenic organisms (e.g., *Pseudomonas aeruginosa*, Figure 4.2.11, right panel) and also to the human enzyme. This example shows the capability of binding site comparison to elucidate similarities between functionally related enzymes which show neither obvious sequence nor structure similarities.

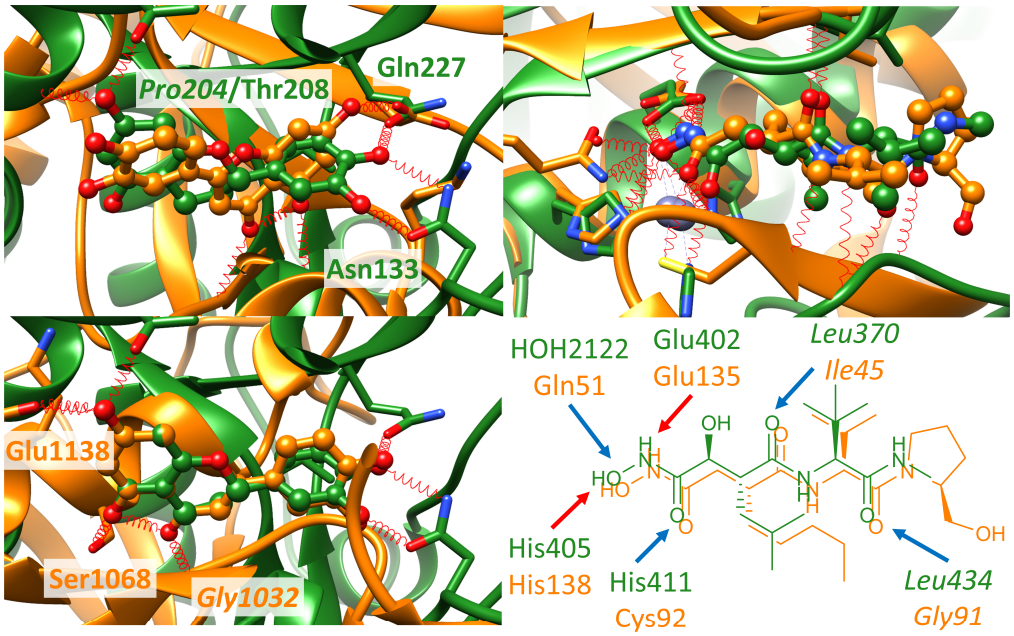


Figure 4.2.11. Examples of dissimilar and similar binding sites bound to highly similar ligands. Left: SiteHopper- (top) and ligand-based (bottom) alignment of the binding sites of tankyrase 2 (PDB-ID 4hkn, ligand-ID LU2) and dihydroflavonol-4-reductase (PDB-ID 2nnl, ligand-ID ERD). Despite their common shape and physicochemical properties, both ligands interact in a unique way with their target. Right: Alignment of the metalloproteases *E. coli* peptide deformylase (PDB-ID 3k6l, ligand-ID 2BB) and human ADAMTS-1 (PDB-ID 2jih, ligand-ID 097) in complex with bound peptide-mimetic compounds (top). The crucial common interactions are depicted separately with the corresponding ligand 2D representations (bottom). Red and blue arrows indicate hydrogen bond acceptor and donor interactions from a target-based view. Residues in italics interact via their backbone atoms.

Taking these examples into account, we can derive two critical issues which have to be considered in binding site comparison studies. First, the researcher must be aware of the score distributions. For tools which do not provide measures to highlight the significance of the found matches, defined score ranges that correspond to similar binding sites have to be considered. Second, a purely shape-based similarity is not sufficient to indicate similar binding sites. Both physicochemical and shape features are necessary to identify meaningful similarities. A dataset that reflects true binding site similarities has to be generated in a more elaborate, but extremely biased manner.

Given the scores obtained with SiteHopper for highly similar binding sites, a threshold was applied to exclude binding sites with low binding site similarities (Figure 4.2.12). This led to a decrease from 15,339 to only 4,487 similar binding site pairs. Intriguingly, tools which were previously characterised by an inferior performance for the initial dataset 5 give promising results for this reduced set, especially with respect to the AUC. Obviously, a predominant shape similarity is characterised by a low score for these tools. SiteAlign, PocketMatch, TM-align, SiteEngine, IsoMIF, and KRIPO perform best besides SiteHopper. However, the good performance of SiteEngine must be attributed to the DistanceScore. In contrast to the CurvatureScore, which was the best choice for the previously discussed datasets and estimates the shape overlap of the matched surface patches, the DistanceScore reflects the spatial similarities between matched pseudo-centres as the sum over all matched centres. The significant performance differences between SiteAlign and SiteEngine with respect to the scoring measures highlight the importance of parameter adjustments for different datasets.

Satisfactory performance can also be shown for nearly all interaction-based methods applied to the trimmed dataset. Nevertheless, the use of similarity thresholds is questionable. The otherwise good performing methods SMAP and ProBiS fail to enrich numerous site pairs which are assigned a high similarity score by SiteHopper. As significant weaknesses of SMAP could not be identified in all previous analyses, it cannot be precluded that the dataset holds further similar binding sites which are declared as “inactives”. However, binding site similarity “lies in the eye of the beholder”. Even purely shape-based similarities might prove helpful for application domains such as drug repurposing or off-target prediction.

Overcoming this thin line between sufficiently high shape similarity, but also a significant physicochemical similarity is a crucial step when analysing binding site comparison tools to the end of drug repositioning or polypharmacology purposes. While the overall similarity has to be sufficiently high to allow for the binding of similar ligands, the binding site properties should also allow for minor dissimilarities as a promising starting point to establish selectivity. Nonetheless, high shape similarities and a sparse number of common hydrogen bond donor and acceptor features of the binding site might hint at potential off-targets. Finally, it can be hypothesised that the meaning of binding site similarities with respect to shape is of interest, but especially repurposing projects and the analysis of common protein-ligand interactions and PPIs will significantly benefit from sufficiently high physicochemical similarities which are readily identified by most of the analysed tools.

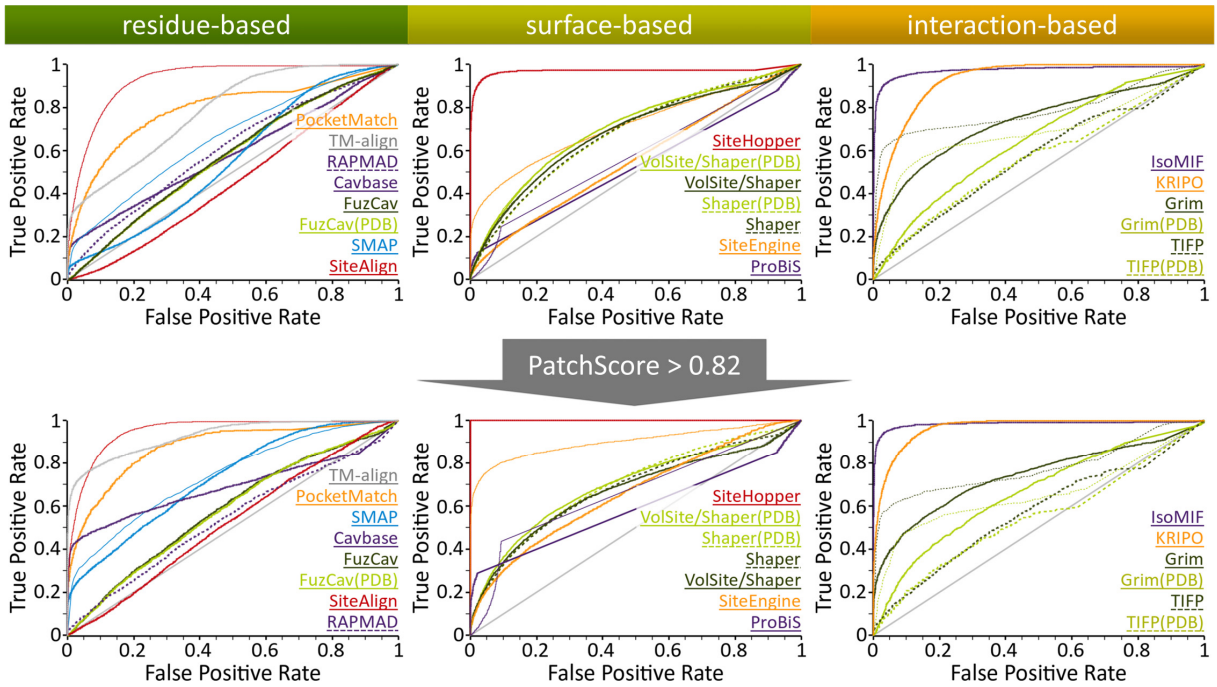


Figure 4.2.12. Evaluation of different binding site comparison tools with two versions of dataset 5. Top: ROC curves for different binding site comparison methods for the initial purely ligand-based dataset. Bottom: ROC curves for the methods if only matches with a promising SiteHopper PatchScore are considered. The name of the tool is coloured according to its corresponding ROC curve. The tools are sorted in descending order with respect to their AUC. The scoring measures that yielded the highest AUC were distance d1 and the Tanimoto coefficient for SiteAlign (thin red line) and SMAP (thin blue line), respectively. SiteEngine showed the highest AUC values when using the similarity DistanceScore (thin orange line). TIFP AUC values significantly improved for the Hamming distance as the distance measure (thin dark green line).

Promiscuous Binding

Finally, potential reasons for promiscuous binding within the sc-PDB-derived dataset were investigated. Studies concerning ligand promiscuity highlighted some crucial small molecule descriptors which might reveal promiscuous binders, e.g., hydrophobicity, number of aromatic rings etc.[390–392] As a matter of fact, we should find complementary properties within the binding sites of proteins in complex with promiscuous ligands. As a basis for this analysis, the "super-promiscuous" ligands assigned by Sturm *et al.*[357] were taken into account. Matches of binding sites in complex with ligands with the PDB names ASD, STR, QUE, DES, and CHD (RTL, RBF, and AE2 were not found as ligands of the active site pairs of dataset 5) are characterised by a low SiteHopper-derived physicochemical and overall similarity. For ASD, binding sites that show a PatchScore above the previously-defined threshold were also identified. Nonetheless, the corresponding alignment shows that both sites are dissimilar with respect to the bound ligands. Consequently, these compounds belong to the group of ligands binding to dissimilar enzymes with distant cavities. Therefore, it was of interest to answer the question of why those ligands bind to different proteins without detectable physicochemical similarities.

Analysing the pocket properties hydrophobicity and polarity by means of DoGSite[254] shows a distinct trend for the corresponding binding sites (Figure 4.2.13, highlighted). Most pockets are characterised by a high hydrophobicity and a low polarity.

An extension of this analysis towards the ligands that were characterised by similar shape and physicochemical properties but bind to dissimilar binding sites led to a similar trend which was also seen for the previously characterised "super-promiscuous" ligands (Figure 4.2.13). Therefore, the question arises whether typical "side-effect" targets[393] can be identified by calculating these easily accessible binding site descriptors. Going one step back and looking for targets in the sc-PDB that bind to the promiscuous molecules in this study further underlines the results. Among them are human serum albumin,[394] PETN reductase,[395] and the mineralocorticoid receptor[396] besides GPCRs, ion channels, and transporters.

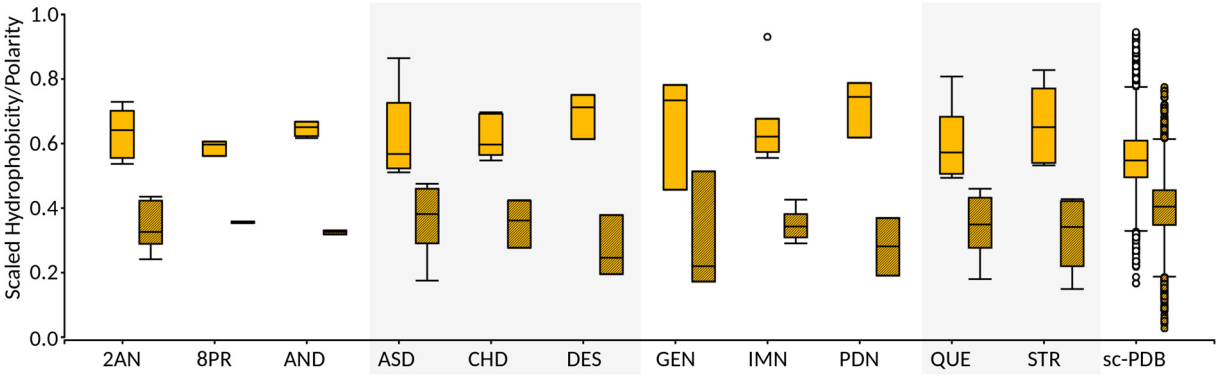


Figure 4.2.13. Hydrophobicity and polarity of dissimilar binding sites in complex with similar ligands. Boxplots of the scaled (min-max normalisation) hydrophobicity (yellow boxes) and polarity (dark yellow boxes) scores of dissimilar binding sites (according to the SiteHopper similarity scores) in complex with ligands that bind to multiple proteins are shown. For comparison purposes, the boxplot for these properties is also given for all ligands of the sc-PDB. "Super-promiscuous" ligands as defined by an earlier study[357] are highlighted in grey.

Although promiscuity is of special importance for the function of the binding sites of certain enzymes, the targeting of promiscuous binding sites in medicinal chemistry should be avoided to lower the risk of potential side-effects. An analysis of binding site descriptors might prove beneficial in approaches trying to evaluate the promiscuity of the targets of interest.[397,398]

Besides those compound types, cofactor-related compounds (e.g., 5GP, AMP, APC, APR, CDP, DG3, DTP, IBM, IMP, RBF, and XMP) and flavone-like molecules (AGI and MYC) were identified to bind to very dissimilar binding sites. In contrast, their corresponding binding site properties followed the opposite trend, i.e., the binding sites are more polar and less hydrophobic. Accordingly, the corresponding molecules are characterised by a high number of hydrogen bond donor and acceptor atoms, negatively charged atoms, and aromatic ring systems. These characteristics enable a multitude of unique and selective interactions with multiple targets. However, the relation of hydrophobicity to polarity for the pockets of the flavone-like compounds AGI and MYC tends towards that of the "super-promiscuous" binders.

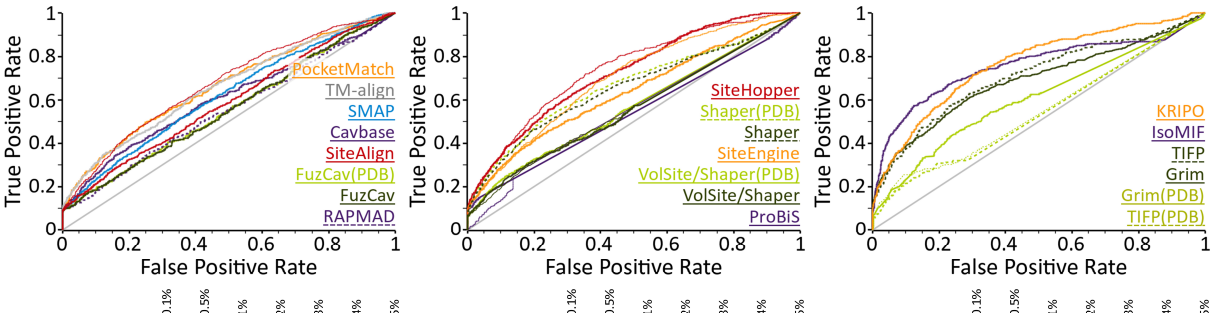
The overall outcome of this analysis shows that we have to adjust our view on promiscuity. Binding sites of ligands that were previously characterised as promiscuous binders are characterised by a high hydrophobicity, a high aromatic ring count, or positive ionisable atoms.[391] Correspondingly, we can find binding site descriptors that highlight protein pockets which are probably able to accommodate a varying number of different, unselectively binding ligands due to their high hydrophobicity.

Therefore, the analysis of promiscuity should not only consider the ligand-based perspective, but should also exploit structural knowledge of protein cavities, comparison tools, and the available pocket descriptors to prevent the targeting of binding sites with promiscuous properties.[391] Particular care has to be taken with respect to druggability. While a high hydrophobicity is an accepted indicator of druggability,[8] a high hydrophobicity might reflect a promiscuous binding site that accommodates varying ligand types in an unselective manner.

The Kahraman Dataset

A commonly analysed dataset for the evaluation of binding site comparison algorithms is the assembly of protein-ligand complexes used by Kahraman and co-workers.[265] It contains several different cofactor and small molecule binding sites (see Table A2.3) and was originally designed to assess the common assumption that ligand and binding site shape are related. It was concluded that differently shaped pockets bind the same ligand, but that shape complementarity might be a significant driver of ligand recognition.

The structures of the dataset were used to estimate the impact of binding site features, interaction patterns, and binding site shape on the recovery of protein pockets binding to identical ligands. The dataset also includes small phosphate binding sites. Various tools failed to process these binding sites due to the small number of interactions or involved binding site residues. To analyse the impact of these binding sites on the outcome, two evaluation steps were conducted: the analysis of 100 vs. 100 structures (Figure A2.8) and the analysis of 80 vs. 80 structures (exclusion of the phosphate binding sites, Figure 4.2.14).



	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	6.5	7.0	7.0	4.9	3.5	3.2	2.9
FuzCav	6.5	7.0	7.0	4.9	3.3	2.8	2.3
FuzCav (PDB)	6.5	7.0	7.0	4.9	3.3	2.9	2.4
PocketMatch	6.5	7.0	7.0	5.3	4.2	3.8	3.6
RAPMAD	6.5	7.0	7.0	4.9	3.5	2.7	2.3
SiteAlign	6.5	7.0	7.0	4.9	3.8	3.1	2.7
SMAP	6.5	7.0	7.0	5.1	3.9	3.1	2.8
TM-align	6.5	7.0	7.0	5.8	4.6	3.9	3.6

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
ProBiS	6.5	7.0	7.0	5.1	3.7	3.3	2.9
Shaper	6.5	7.0	7.0	5.1	4.3	3.9	3.5
Shaper (PDB)	6.5	7.0	7.0	5.2	4.5	4.0	3.7
SiteEngine	6.5	7.0	7.0	5.4	4.3	3.6	3.2
SiteHopper	6.5	7.0	7.0	5.8	4.8	4.2	3.8
VolSite/Shaper	6.5	7.0	6.5	4.1	3.2	2.7	2.4
VolSite/Shaper (PDB)	6.5	7.0	6.1	4.0	3.1	2.8	2.5

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Grim	6.5	7.0	6.2	5.9	5.4	4.7	4.4
Grim (PDB)	0.0	2.4	4.7	3.8	3.4	2.7	2.1
IsoMIF	6.5	7.0	7.0	5.4	5.2	5.1	4.8
KRIPO	6.5	7.0	7.0	5.9	5.1	4.6	4.2
TIFP	6.5	7.0	7.0	6.0	5.0	4.3	4.0
TIFP (PDB)	0.0	3.9	5.3	3.3	2.6	2.5	2.5

Figure 4.2.14. Evaluation of different binding site comparison tools with respect to dataset 6. Top: ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is coloured according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. The best AUC for SiteAlign resulted from the distance measure d1 (thin red line). For ProBiS, VolSite/Shaper, SiteEngine, and SiteHopper the scores SVA, Tanimoto (color), TotalScore, and ShapeTanimoto yielded the best AUC values (thin lines). For TIFP(PDB), the use of the Hamming distance led to the best results with respect to the AUC (thin line). Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

The trend that was already recognisable for the ROCS analysis-derived binding site pairs becomes even more pronounced for this dataset. SiteHopper, IsoMIF, and KRIPO significantly outperform the other tools. This tendency is less obvious if the phosphate binding sites are omitted from the analysis. This suggests that even similarities between the small phosphate binding sites are recognised by the tools that performed best for this dataset. While the performance of residue-based tools is poor, that of surface- and interaction-based methods looks promising. The interaction-based method KRIPO outperforms all residue- and surface-based methods except SiteHopper and shows a high early enrichment.

The early EFs of IsoMIF and KRIPO are the highest for all interaction-based methods if all 100 structures are considered (dataset 6.2, see Figure A2.8). IsoMIF and SiteHopper are the tools with the highest AUC for this complete set. For IsoMIF, this is expectable as the method was evaluated on and probably optimised for these binding site pairs.

A preliminary conclusion that can be drawn from both analyses is that a combination of the surface- and interaction-based methods might be key for the detection of binding site similarities between otherwise unrelated proteins. The similarity matrix of the all-against-all comparisons with IsoMIF and SiteHopper underlines this finding (Figure 4.2.15). Both methods successfully differentiate between similar and dissimilar pairs and they are even able to cluster binding sites according to the bound ligands. However, their overall poor performance suggests that some similarities are neglected and that there are also local similarities which are not considered in the dataset. While SiteHopper assigns high scores to the local binding site similarities between sites that contain phosphate and the phosphate-containing ligands AMP, AND, and ATP, the KRIPO scores are considerably low in these cases. However, the noise in the matrix is higher for SiteHopper.

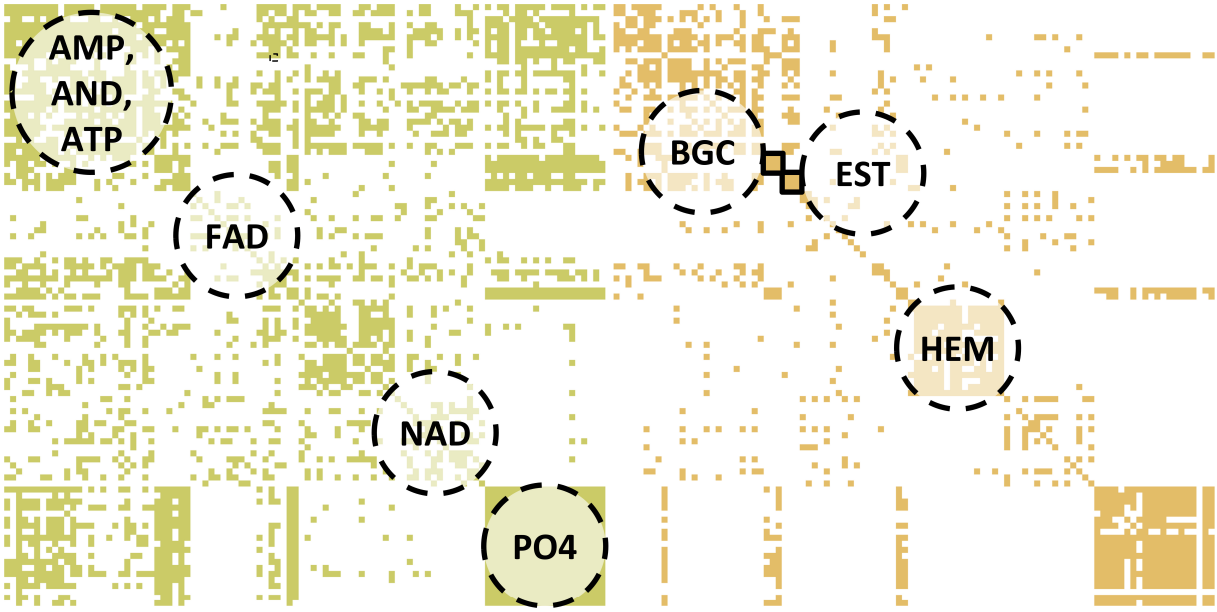
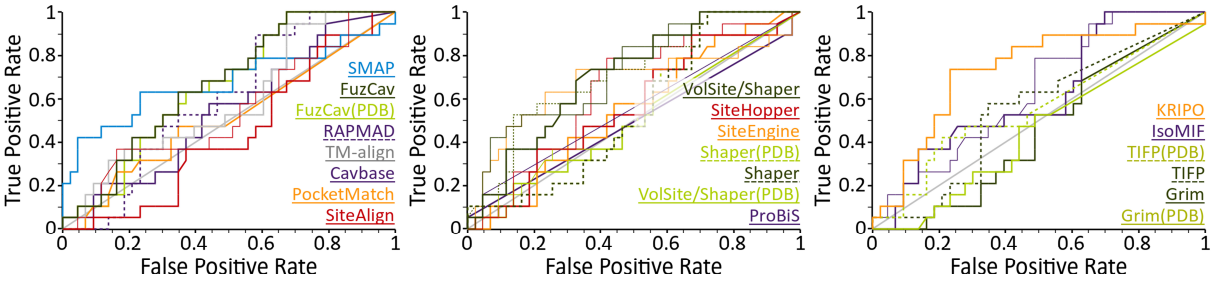


Figure 4.2.15. Binding site similarities in dataset 6.2. The similarity score matrices for the Kahraman dataset generated from the SiteHopper (left) and IsoMIF (right) comparisons are presented. Both methods find clusters of similar binding sites with identical ligands. The combination of both methods might even lead to improved differentiation. SiteHopper similarity scores (PatchScore) above 0.65 are coloured green. IsoMIF similarity scores (tani) above 0.4 are coloured yellow.

The Barelier Dataset

A recent publication focuses on the similarities and differences between binding sites of unrelated proteins binding to similar ligands. The authors screened the PDB for identical ligands, but they performed some pre-filtering steps (ligands found in at least 15 complexes, with a molecular weight above 500 g/mol, or with less than 10 non-hydrogen atoms, as well as cofactors were excluded). The resulting 62 pairs of protein-ligand complexes were classified according to the similarities of the ligand interactions with the cavity residues. The authors divided the dataset into three classes: similar interactions of the ligands with similar protein functionalities (class A), similar interactions of the ligands with different functional groups of both proteins (class B), and different functional groups of the ligands interacting with different sub-pockets in the proteins (class C). The calculation of the vdW and electrostatic scores per atom and the normalisation of both scores enabled this detailed comparison of the ligand interactions in the pairs of complexes.[267]

The protein-ligand complexes of class A were used as structures with similar binding sites in this analysis. Due to the sparse number of protein pairs, EFs at 1.6%, 8.1%, 16.1%, 32.3%, 48.4%, 64.5%, and 80.6% of the ranked dataset were calculated. The results for all methods are presented in Figure 4.2.16. The ROC curves show a nearly random ranking of similar and dissimilar binding site pairs for the methods analysed herein. Additionally, the improvement in the AUC values using the most appropriate scoring measures is negligible for this dataset.



	$EF_{1.6\%}$	$EF_{8.1\%}$	$EF_{16.1\%}$	$EF_{32.2\%}$	$EF_{48.4\%}$	$EF_{64.5\%}$	$EF_{80.6\%}$
Cavbase	3.3	0.7	1.3	0.8	1.1	1.0	1.0
FuzCav	3.3	1.3	1.0	1.3	1.3	1.2	1.2
FuzCav (PDB)	3.3	1.3	1.0	1.3	1.3	1.1	1.2
PocketMatch	0.0	0.7	1.0	1.0	1.0	0.7	0.6
RAPMAD	0.0	0.0	0.3	1.1	1.1	1.1	1.2
SiteAlign	0.0	0.0	0.3	0.7	0.8	1.0	1.0
SMAP	3.3	2.6	2.3	1.6	1.3	1.1	1.0
TM-align	3.3	0.7	1.3	1.1	1.0	1.1	1.2

	$EF_{1.6\%}$	$EF_{8.1\%}$	$EF_{16.1\%}$	$EF_{32.2\%}$	$EF_{48.4\%}$	$EF_{64.5\%}$	$EF_{80.6\%}$
ProBiS	3.3	0.7	0.3	0.2	0.1	0.1	0.5
Shaper	3.3	1.3	0.7	0.7	1.0	1.1	1.2
Shaper (PDB)	3.3	1.3	0.7	0.7	1.0	1.1	1.2
SiteEngine	0.0	0.7	0.7	1.0	1.1	1.1	1.1
SiteHopper	0.0	0.7	0.7	1.1	1.1	1.1	1.1
VolSite/Shaper	3.3	1.3	1.3	1.5	1.4	1.2	1.2
VolSite/Shaper (PDB)	0.0	0.7	1.0	1.0	1.0	0.9	0.7

	$EF_{1.6\%}$	$EF_{8.1\%}$	$EF_{16.1\%}$	$EF_{32.2\%}$	$EF_{48.4\%}$	$EF_{64.5\%}$	$EF_{80.6\%}$
Grim	0.0	0.0	0.3	0.7	0.9	1.0	0.8
Grim (PDB)	0.0	0.0	0.3	0.8	1.0	0.8	0.7
IsoMIF	0.0	0.7	1.3	1.3	1.1	1.1	1.2
KRIPO	0.0	0.0	0.0	2.6	1.8	1.3	2.1
TIFP	3.3	0.7	0.3	0.8	1.2	1.1	0.8
TIFP (PDB)	3.3	0.7	1.0	1.3	1.0	0.8	0.7

Figure 4.2.16. Evaluation of different binding site comparison tools with respect to dataset 7. Top: ROC curves for residue- (left), surface- (right), and interaction-based (centre) comparison methods. The name of the tool is coloured according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. The thin red line represents the resulting ROC curve for SiteAlign for the distance score d1. Thin lines represent the ROC curves for ProBiS, Shaper, Shaper(PDB), VolSite/Shaper, VolSite/Shaper(PDB), SiteEngine and SiteHopper if the scoring schemes SVA, FitTversky (color), FitTversky (color), RefTversky (color), Tanimoto (fit), DistanceScore, and ShapeTanimoto are used, respectively. The thin purple line represents the resulting ROC curve for IsoMIF and the taniMW similarity score. Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

The decision for suitable tools based on the ROC curves is not possible as nearly all differences are insignificant (see Table A2.24). Additionally, the dataset is too small to pinpoint promising tools. In comparison to all earlier datasets, the substandard performance of the methods is unique. Therefore, the underlying binding site alignments of class A pairs were investigated for the tools that showed a comparatively high early enrichment and allow the visualisation of the underlying matches (Cavbase, SMAP, TM-align, ProBiS, and VolSite/Shaper). Based on the per-atom score analysis of Barelier *et al.*,[267] it can be expected that the corresponding ligand atoms nicely overlay for high scoring pairs of binding sites. As can be learned from Figure 4.2.17, a visual inspection of the obtained binding site alignment is often inevitable to analyse the validity of found high-scoring site pairs.

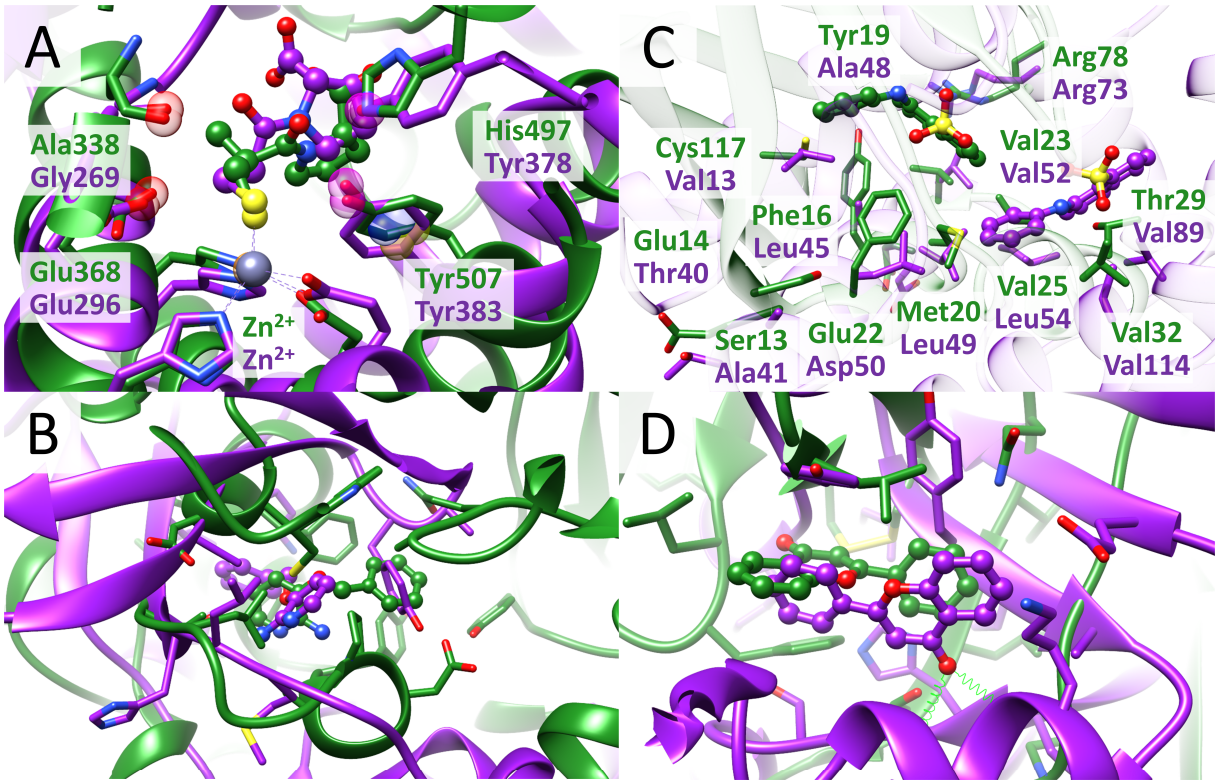


Figure 4.2.17. The impact of alignment visualisation on binding site comparison. Alignments of high-scoring binding site pairs of the Barelier dataset generated by Cavbase (A), TM-align (B), SMAP (C), and Shaper (D) are presented. (A) Superposition of angiotensin-converting enzyme (PDB-ID 2x8z, green) and leukotriene A4 hydrolase (PDB-ID 4dpr, purple) in complex with captopril (ball-and-stick). Red spheres denote hydrogen bond acceptor features, while purple spheres represent mixed hydrogen bond acceptor/donor features. Metal coordination sites are marked by orange spheres. Blue and yellow spheres denote residues with aromatic and aliphatic characteristics. The Cavbase similarity score for this match is 11.37. (B) Alignment of leukotriene A4 hydrolase (PDB-ID 3fty, green) and mitogen-activated protein kinase 14 (PDB-ID 1w7h, purple) crystallized with the small molecule fragment 3-phenylmethoxypyridin-2-amine (ball-and-stick). Residues within a 4 Å radius of any ligand atom are depicted in stick representation. This alignment yields a TM-score of 0.32. (C) Superposition of adipocyte lipid-binding protein (PDB-ID 2ans, green) and pheromone-binding protein (PDB-ID 1ow4, purple) in complex with the fluorescent probe 8-anilino-1-naphthalene sulfonate (ball-and-stick). The residues shown in stick representation represent only a fraction of all matched residues. The SMAP Raw Score for this site pair is 63.44. (D) Shaper-based alignment of the neocarzinostatin (PDB-ID 2g0l, green) and tankyrase 2 (PDB-ID 4hki, purple) flavone (ball-and-stick) sites (TanimotoCombo = 0.92). Residues within a 4 Å radius of the ligand are represented as sticks. Hydrogen bond interactions are depicted as green springs.

The best scored Cavbase match reflects an acceptable overlay of captopril in complex with angiotensin-converting enzyme and leukotriene A4 hydrolase (Figure 4.2.17 (A)). The match between these binding sites was also the highest scoring one within all similar pairs for TM-align. A good superposition of the bound identical ligands could be achieved. However, the second highest ranked true positive match yielded a superposition with poorly aligned ligands (Figure 4.2.17 (B)). Similarly, the SMAP alignment of the class A pair with the highest rank does not superpose the corresponding ligand atoms as expected. Nevertheless, it is obvious that the tool was able to find similarities between the residues of both binding sites. A closer look shows that the residues considered for the comparison are broadly distributed across the proteins (Figure 4.2.17 (C)) providing a plausible explanation for the inadequate alignment with respect to the ligands. Only one significant (according to the e-value) match was found by ProBiS for the binding sites of the proteins with the PDB-IDs 2x8z and 4dpr. The ligands align well for the corresponding structural superposition. Shaper-derived alignments also do not provide perfect alignments of ligand atoms for the best scored hits. Nonetheless, ligand moieties with a similar physicochemical nature were superimposed appropriately. These observations highlight the importance of alignment visualisations for the estimation of the validity of the binding site similarities. As introduced earlier for dataset 5, Sturm *et al.*[357] analysed the chemical nature of promiscuous ligands. Based on an analysis of the sc-PDB the authors named the quinone with the PDB ligand name QUE as a super-promiscuous ligand. As presented in Figure 4.2.17 (D), each ligand interacts mainly via hydrophobic interactions with the residues in the cavity. In the binding site of the structure with the PDB-ID 4hki, the ligand is engaged in two hydrogen bonding interactions with the protein backbone. Upon these observations, the question arises whether a binding site comparison of predominantly lipophilic pockets will reveal interesting starting points for repurposing projects.

As cofactors and ions were considered for the comparisons of the interactions by Barelier *et al.*,[267] they were included to assess their impact on the comparison outcome (dataset 5.2). Although some tools are not suited to consider non-protein molecules as part of the binding site, this ensures that none of the methods is significantly influenced by the presence or absence of other ligands. The outcome did not lead to significant changes in the obtained results (see Figure A2.9 in Appendix 2). Although some slight improvements in the AUC could be observed, none of the tools ranked the class A cavity pairs considerably higher than the dissimilar binding site pairs of classes B and C.

According to these results, the only acceptable match found within all class A pairs was that between the captopril binding sites of the proteins with the PDB-IDs 2x8z and 4dpr. Although some binding site similarities of class A pairs resulted in a high score, the alignment of the binding sites did not correctly superpose the counterpart ligand atoms but some physicochemically similar ligand moieties (Figure 4.2.17 (D)). A plausible explanation is the fact that identical ligand atoms “observe” similar interaction partners, but the conformation of both molecules is very different (see Table A2.4 for the RMSD values of the ligand conformations in the sites of all cavity pairs). However, this is not the only explanation for the overall poor results for all comparison methods.

The question arose whether different parts of the binding sites are not sufficiently dissimilar to be clearly distinguished (promiscuous ligand binding sites). The binding site analysis of the structures involved in the different classes of pairs and the sites of a sequence-culled subset of the sc-PDB[10] of druggable binding sites led to the results presented in Figure 4.2.18. While the hydrophobicity and aromaticity of the binding sites in the set of Barelier *et al.* are higher than that for the sc-PDB structures, the number of hydrogen bond donor and acceptor atoms is comparatively lower. This underlines the finding that binding site similarity does not always result in a similar binding mode of the bound ligands. For hydrophobic and aromatic ligands, multiple binding modes are possible irrespective of the underlying binding site alignment. Moreover, the generation of an accurate binding site alignment is hampered by the fact that a broad variety of acceptable cavity alignments are found for highly hydrophobic binding sites. It can be concluded that the generation of a similar dataset considering only druggable binding sites with a broader range of site features was more valuable to evaluate the performance of binding site comparison methods.

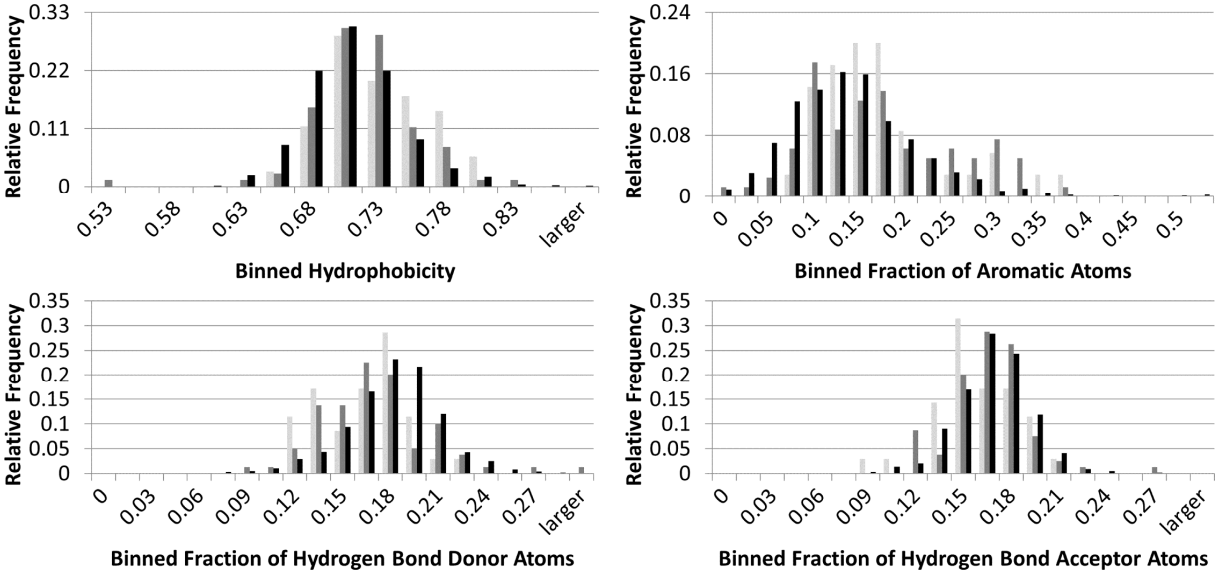
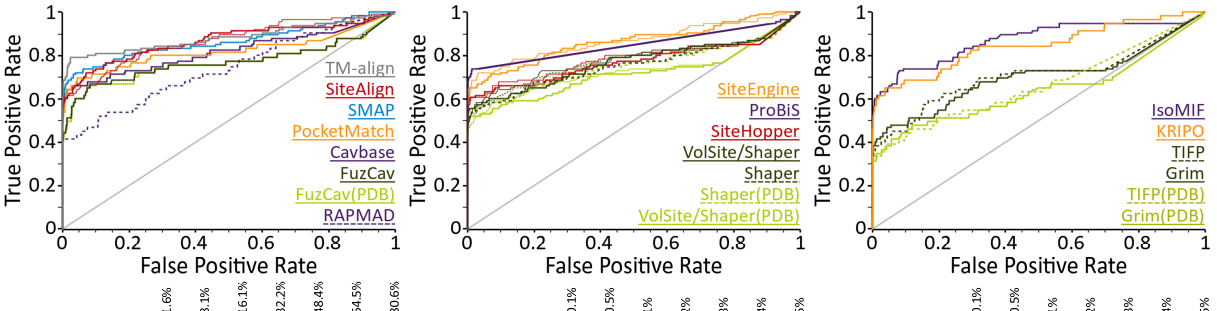


Figure 4.2.18. Properties of the binding sites of dataset 7. Results of the binding site feature analysis for the class A, B, and C site pairs of Barelier *et al.*[267] and a sequence-culled subset of druggable sites are given. The relative frequencies of the binned properties are presented in light grey for class A structures, in dark grey for structures belonging to class B and C pairs, and in black for the sc-PDB[10] subset. The binding site features were calculated with DoGSite.[254]

The Dataset of Successful Examples

The final dataset of this benchmark study is composed of known binding site similarities that are summarised in the literature.[48] Cavities with known similarities (see Table A2.5 for an overview) were integrated into a sequence-culled sc-PDB subset. Their recovery within the highest ranked pairs was analysed. As compared to the previous datasets, the “active” pairs represent practically relevant binding site similarities ranging from obvious similarities to related sites in structurally distant proteins. They should be identified and respectively scored by all comparison methods. Figure 4.2.19 shows the outcome of this analysis and puts the results obtained for the dataset of binding sites with similar ligands (dataset 5) and dataset 6 into perspective.

The differences between the diverse types of binding site comparison tools cancel each other out for this dataset. There are only minor differences in the AUC values for the methods and most of these differences are insignificant (see Table A2.27 for the significant differences). The promising results of this analysis indicate the general applicability of all methods analysed herein. However, we observe some trends which are worth mentioning.



	$EF_{1.6\%}$	$EF_{8.1\%}$	$EF_{16.1\%}$	$EF_{32.2\%}$	$EF_{48.4\%}$	$EF_{64.5\%}$	$EF_{80.6\%}$
Cavbase	339.1	118.3	62.6	32.2	21.4	16.5	13.2
FuzCav	408.7	88.7	44.3	25.7	18.8	15.0	12.0
FuzCav (PDB)	408.7	85.2	44.3	24.8	18.3	15.0	12.0
PocketMatch	434.8	104.3	59.1	32.2	22.0	16.5	13.9
RAPMAD	400.0	83.5	41.7	20.9	13.9	10.4	8.7
SiteAlign	426.1	116.5	60.0	30.9	20.9	16.5	13.4
SMAP	313.0	118.3	65.2	34.3	23.2	17.8	14.4
TM-align	434.8	135.7	68.7	37.8	26.4	19.8	15.8

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
ProBiS	382.6	137.4	70.4	37.0	24.6	18.5	14.8
Shaper	417.4	99.1	53.0	27.0	18.3	13.9	11.3
Shaper (PDB)	417.4	100.9	53.0	27.0	18.8	14.1	11.3
SiteEngine	373.9	118.3	65.2	32.6	22.0	16.7	13.6
SiteHopper	400.0	114.8	60.0	30.4	20.6	15.4	12.3
VolSite/Shaper	408.7	107.8	55.7	27.8	18.8	14.3	11.7
VolSite/Shaper (PDB)	400.0	102.6	52.2	26.1	17.4	13.0	10.6

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Grim	313.0	76.5	40.0	20.9	14.2	11.1	9.4
Grim (PDB)	52.2	62.6	31.3	17.4	12.2	9.1	7.7
IsoMIF	408.7	107.8	60.0	30.9	21.2	15.9	13.0
KRIPO	426.1	106.1	56.5	30.9	20.6	16.3	13.0
TIFP	330.4	69.6	36.5	19.1	13.3	10.4	8.7
TIFP (PDB)	0.0	64.3	33.9	17.0	11.9	9.3	7.8

Figure 4.2.19. Evaluation of binding site comparison tools with respect to dataset 8. Top: The ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is coloured according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. SiteAlign yielded a slightly better AUC if the distance d1 was used (thin line). The best AUC values for ProBiS, Shaper, Shaper(PDB), VolSite/Shaper, VolSite/Shaper(PDB), SiteEngine, and SiteHopper resulted from the scoring measures z-score, Tanimoto (color), Tanimoto (color), Tanimoto (color), Tanimoto (color), TotalScore, and ShapeTanimoto, respectively (thin lines). Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

Most residue-based methods outperform the surface- and interaction-based approaches. The latter tools were the best among all for the dataset of Kahraman and co-workers. TM-align – a tool which cannot be used to score physicochemical binding site differences and aligns residues exclusively based on geometric criteria – is the one with the highest AUC and the highest early enrichment. This underlines a specific challenge for binding site comparison. Although the software should be able to accurately score similarities, it also must allow for the detection of non-obvious similarities. Additionally, the difference in performance of Cavbase and RAPMAD is very pronounced for this set of realistic similar protein pairs. Obviously, a purely histogram-based comparison enriches dissimilar pairs together with the similar ones.

Surface- and interaction-based tools are also characterised by a high early enrichment of known similar pairs. SiteEngine and ProBiS are the surface-based methods which are best suited to recover known similarities in this set. Despite the enormous difference in their ability to score site similarities according to the number of substituted residues (datasets 3 and 4), both tools perform similarly in this analysis underlining the necessity to choose a binding site comparison tool based on the user requirements.

Regarding the interaction-based methods, the superior performance of IsoMIF and KRIPO that was present for most other datasets becomes even more evident here. This trend might change upon the exclusion of binding sites that could not be processed for a comparison with Grim and TIFP which perform equally good for this set. Although all interaction-based methods show a comparable early enrichment, KRIPO and IsoMIF seem to be the methods of choice for applications which originate from already known protein-ligand complexes. This conforms with the finding that the worse performing methods Grim and TIFP do not identify related sites in complex with different ligands, i.e. small molecules with ECFP4-based pairwise Tanimoto coefficients below 0.6.

This analysis was not designed for the identification of the best method but emphasises the impact of the study design on the final decision. For example, while ProBiS can be safely applied to perform query-based searches for similar binding sites (e.g., for function prediction), SiteEngine should be chosen to perform an all-against-all comparison between binding sites of similar proteins. Due to its superior similarity scoring performance, it can be used to elucidate site relationships and name potential off-targets.

4.2.6 Can We Distinguish Between Similar and Dissimilar Sites?

Similarity Thresholds

Another topic of interest, which was already outlined with respect to dataset 1, is the definition of appropriate score cut-offs to differentiate between similar and dissimilar binding sites. The problems and dataset dependence of similarity cut-off values were exhaustively discussed for small molecule similarities.[399–401] These analyses in cheminformatics are supported by the availability of bioactivity data.[402,403] In consequence, some fingerprint similarity thresholds to distinguish between similar and dissimilar molecules were derived.[400]

In contrast, bioactivity annotations are hard to extract for small molecule binding sites. The binding site of molecules with known bioactivities is often difficult to spot in bioassays. Especially enzyme activity assays provide little evidence for the modulator's binding site. For most known bioactive small molecules, only the respective target is reported in publicly available databases such as ChEMBL.[303] The definition of the similarity of binding sites relies on structural data retrieved from the PDB[7] and comparative modelling studies. The investigations of Barelier and co-workers[267] underline the sparseness of this crucial knowledge resource. The evaluation of binding site similarities is therefore restricted to well-known binding sites (e.g., cofactor binding sites), artificially created dataset (e.g., the decoy datasets applied in this study), and already known binding site similarities. These factors impede the evaluation of the available comparison methods and the definition of a reliable threshold per method. In contrast to small molecules, the unique definition of binding site dimensions, properties, and locations is difficult if not infeasible. Reasons are the binding site flexibility, the presence of sub-pockets, and the dependence of the binding site boundaries on the ligand. Additionally, the problem of deriving reliable cut-off values for classification purposes is still up for discussion. Youden's index,[404] the F-measure,[405] likelihoods,[406] or discriminant power[407] are only a few examples for discriminant measures.

The ROC-based Youden's J statistic[404] was employed to find optimum cut-off scores to distinguish between similar and dissimilar site pairs for each method. First, only the tools' results for dataset 1 was used to define a unique threshold per method. The application of these thresholds for the dataset of successful applications (dataset 8) leads to the results presented in the first part of Table 4.2.5. While all tools except RAPMAD are able to reliably reject dissimilar binding site pairs with the given cut-offs (high specificity), the maximum sensitivity is 0.73. A considerable number of similar binding site pairs yield similarity scores below the predefined thresholds and are therefore rejected. In practice, that means that similar binding sites are neglected due to the predefined cut-off criteria. Astonishingly, PocketMatch and RAPMAD – both tools which showed poor performance for nearly all datasets of this study – are within the three methods with the highest sensitivity for the complete dataset.

Additionally, the developer-assigned thresholds were used. This was possible for FuzCav, SiteAlign (for the distances d1 and d2), Shaper and VolSite/Shaper (for the RefTversky (color) similarity), Grim, and TIFP (for the Tanimoto coefficient). The thresholds for SiteAlign were initially derived by analysing the score distributions of similar and dissimilar sites. In all other cases, the F-measure[405] was applied to identify an optimum cut-off. Using the pre-defined thresholds as recommended by the developers, we observe a worse specificity for FuzCav, SiteAlign, and VolSite/Shaper, but a better sensitivity with respect to dataset 8 as compared to the thresholds defined by applying the AUC-based Youden's J statistic to the results for dataset 1. Depending on the aim of a binding site comparison study, the optimum threshold will vary. For a drug repurposing study, high specificity is crucial because it is interesting to retrieve only highly similar sites. In contrast, high sensitivity values are more suitable for off-target predictions as it is important to find as many potential off-targets as possible.

The optimum cut-off values were also calculated for dataset 8 and are presented the second part of Table 4.2.5. The use of these thresholds leads to better specificity and sensitivity values underpinning the strong dataset dependence of similarity score thresholds. Especially the ability to detect similar binding sites increases. The high specificity, which was already nearly perfect for the initial cut-offs, persists. This observation is most prominent for SiteAlign, TIFP, and IsoMIF. Importantly, not only the sensitivity values, but also the ranking of the methods according to their sensitivity differ for both applied thresholds.

This finding is important for designing a binding site comparison study. Whenever searching for similar binding sites, it is not recommended to reject similarities solely based on the low similarity score. The basic idea is to analyse the resulting list of ranked binding site similarities, visualise high scored pairwise alignments, and finally decide whether the sites are sufficiently similar. The choice for an appropriate tool should always consider the methods' EFs for query-based evaluations. In the case of the analysis of relationships between multiple binding sites or the prediction of off-targets, the method should be chosen based on the AUC values. If a similarity score cut-off for classification purposes is indispensable, the user has to take into account the aim of the comparison as well as the nature of the dataset.

Table 4.2.5. The impact of binding site similarity cut-offs on the comparison outcome. Sensitivity (se) and specificity (sp) values for dataset 8 using different score cut-offs are given. The score thresholds to discriminate similar and dissimilar site pairs were determined using the ROC-based Youden's J statistic[404] based on datasets 1 and 8. Both thresholds were applied for the sensitivity and specificity calculations. The rank column gives the rank of the respective method within all methods for the corresponding sensitivity. Sensitivity and specificity values for thresholds that were assigned by the method developers are given in parentheses. For SiteAlign, Shaper, VolSite/Shaper, and TIFP, the corresponding scoring measure is different from the one used in this study.

Method	**Score Cut-Off Based on Dataset 1** (Cut-Off Provided by the Developers)				**Score Cut-Off Based on Dataset 8**			
	cut-off	***sp***	***se***	***rank***	***cut-off***	***sp***	***se***	***rank***
PocketMatch	0.16	0.86	0.73	1	0.26	0.95	0.70	6
SMAP	69.18	0.97	0.71	2	67.61	0.96	0.72	4
RAPMAD	0.89	0.51	0.71	3	0.97	1.00	0.42	19
TM-align	0.49	0.99	0.70	4	0.38	0.98	0.79	1
KRIPO	0.50	0.88	0.69	5	0.56	0.96	0.65	11
ProBiS	8.03	1.00	0.67	6	2.98	0.99	0.74	2
SiteEngine	875.25	0.98	0.65	7	1,063.70	0.99	0.65	10
IsoMIF	0.49	0.96	0.63	8	0.46	0.92	0.73	3
Cavbase	15.85	0.99	0.63	9	8.33	0.97	0.66	9
SiteAlign	0.85 (d1 < 0.6, d2 < 0.2)	0.99 (0.81, 0.93)	0.60 (0.77, 0.70)	10	0.80	0.93	0.72	5
FuzCav (PDB)	0.21 (0.16)	0.94 (0.70)	0.60 (0.74)	11	0.20	0.92	0.67	7
FuzCav	0.21 (0.16)	0.94 (0.70)	0.60 (0.74)	12	0.20	0.92	0.67	8
VolSite/ Shaper	0.89 (RefTversky (color) > 0.35)	0.95 (0.59)	0.58 (0.77)	13	0.99	0.99	0.56	15
SiteHopper	0.95	0.99	0.57	14	0.87	0.99	0.61	12
Shaper	0.89 (RefTversky (color) > 0.35)	0.94 (0.52)	0.57 (0.76)	15	0.98	0.99	0.54	16
Shaper (PDB)	0.89 (RefTversky (color) > 0.35)	0.94 (0.51)	0.57 (0.76)	16	0.94	0.97	0.57	14
VolSite/Shaper (PDB)	0.89 (RefTversky (color) > 0.35)	0.96 (0.64)	0.53 (0.68)	17	0.98	0.99	0.52	17
Grim (PDB)	0.57 (0.594)	0.78 (0.96)	0.51 (0.38)	18	0.59	0.94	0.42	20
TIFP	0.23 (Tanimoto > 0.318)	0.89 (0.89)	0.49 (0.49)	19	0.17	0.83	0.59	13
Grim	0.57 (0.594)	0.90 (0.99)	0.48 (0.42)	20	0.58	0.95	0.47	18
TIFP (PDB)	0.20 (Tanimoto > 0.318)	0.89 (0.92)	0.44 (0.41)	21	0.35	0.96	0.39	21

The cut-off values for all benchmark datasets as determined with the method of Youden[404] are given in Table A2.31 in Appendix 2, but they should never be overestimated. The thresholds for dataset 5 were excluded as it was shown that the sites binding to similar ligands in a comparable conformation do not necessarily suggest specific binding site similarities. The variation in the thresholds calculated for the seven datasets shows that cut-off scores have to be carefully chosen dependent on the anticipated outcome. The similarity scoring for binding sites is an ambitious undertaking (see benchmark studies for dataset 1). The reliability of the similarity measures varies with the binding site flexibility, the bound ligands, and the binding site definition. The development of meaningful benchmark datasets for a reliable cut-off assessment is challenging due to the lack of knowledge of site-specific bioactivities and the sparseness of known protein-ligand complexes with similar ligands. Only the growth of this information will enable the establishment of datasets for threshold optimisation.

Meaningful cut-offs cannot be determined before even more protein-ligand complex structures and bioactivity data with binding site annotations are available. In line with other investigations concerning this topic,[53,69,266,408] it is reasoned that these factors preclude the assignment of thresholds for the binary classification of similar and dissimilar site pairs.

The Discovery of Unexplored Site Similarities

Finally, the question arose: “What are the similarities which most tools miss?”. To this end, the boxplots of the similarity score distributions for all methods (see Figures A2.3-5) were taken as a basis for the identification of “challenging cavity pairs”. Active pairs with similarity or distance values in the range of the dissimilar pairs for most methods were analysed for dataset 8. The methods with the highest scores for these matches were applied to generate the corresponding binding site alignments. Two alignments of challenging binding site matches are depicted in Figure 4.2.20. The general outcome of this analysis reveals the difficulties in finding these similar pockets. The first example was identified by the residue-based method Cavbase.[356] Intriguingly, a low Cavbase similarity score of 3.25 was assigned for this site pair in the present study. In contrast, the interaction-based method IsoMIF assigned a high score and provided the presented site alignment. It is obvious that this match cannot be found by residue-based methods as the chemical nature of most aligned residues is different. Nevertheless, similarities were detected with respect to putative interaction patterns. SiteAlign was applied for the identification of the match between synapsin and PIM1 kinase.[370] In contrast to the previous example, SiteAlign assigns a high score to the match and it is found within the best scored pairs. A high ranking for this cavity pair was also observed for the surface-based method SiteEngine. The alignment in Figure 4.2.20 illustrates why most methods do not detect this similarity. Although the atoms of both ligands overlay as expected, it is difficult to pinpoint specific residue correspondences. Only sub-parts of both pockets are similar. This final analysis underlines that it is impossible to identify a “best-performing tool”. Some of the similar cavity pairs were within the best scored hits for one, but not all tools. The results of two or more tools are often complementary. This holds especially true for binding sites sharing common interaction patterns or surface properties which are not evident on a residue-based level.

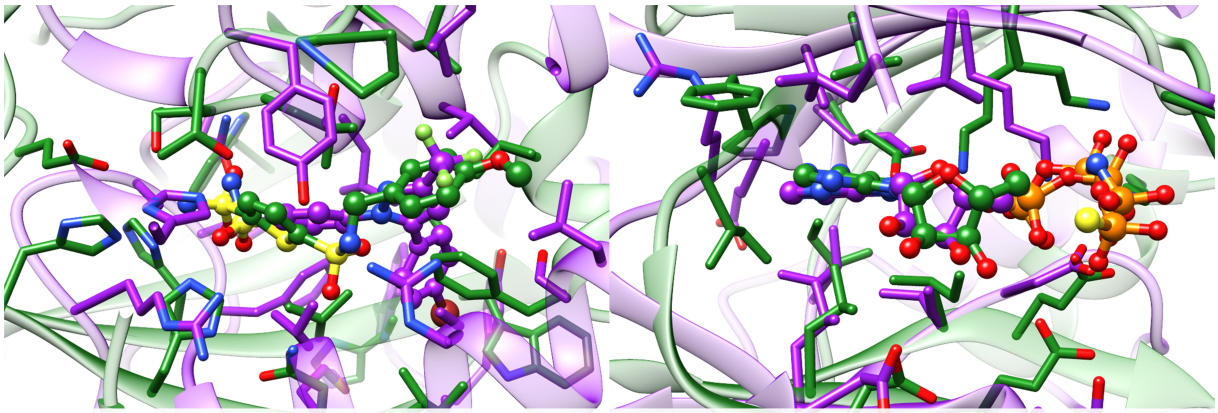


Figure 4.2.20. Binding site similarities which most methods miss. The alignment of human carbonic anhydrase II (PDB-ID 1bn4, green) and cyclooxygenase-2 (PDB-ID 6cox, purple) as obtained with IsoMIF is shown on the left. The binding sites of synapsin (PDB-ID 1aux, green) and PIM1 kinase (PDB-ID 3a99, purple) as aligned by SiteEngine are presented on the right. All illustrations were generated using UCSF Chimera.[228]

4.2.7 Runtime Analyses

Aside from their performance on the benchmark datasets, the comparison algorithms differ in terms of the error rates and computational time. Table 4.2.6 summarises the algorithms in ascending order with respect to their runtime and gives an overview of the computational cost for an all-against-all comparison of the binding sites of dataset 6 (10,000 comparisons). The binding site pre-processing and comparison were evaluated independently. As PocketMatch, SiteAlign, and TM-align rely on the definition of the cavities by the user, the preparation times for the Python-based pre-processing steps are given.

Pocket preparation or comparison failed with some methods (see Figure A2.10 for the failure rates of all tools). The failure rates of TIFP and Grim underline the importance of the PDB to MOL2 conversion to perform the comparisons for most of the dataset structures. The calculation of the average runtimes per comparison takes the number of omitted comparisons into account. For Cavbase and RAPMAD, all binding sites could be processed and prepared, but the ligand-defined binding site was not predicted for nine structures. For FuzCav and Shaper, only 96 binding sites could be prepared. The processing of the prepared pockets (with the pdbconv tool of IChem) with VolSite for a subsequent Shaper comparison (VolSite/Shaper) led to only 57 pockets derived from PDB files and 76 cavities if MOL2 files were used as input. TIFP preparations led to 47 fingerprints for the structures provided in the PDB format and 77 for the MOL2-formatted structures. In contrast, Grim was able to prepare 95 and 96 pockets for MOL2 and PDB files, respectively.

Although a clear correlation between comparison method and runtime cannot be derived from Table 4.2.6, we can differentiate between fast methods (several microseconds per comparison), moderately fast methods (several milliseconds per comparison), and comparatively slow methods (several seconds per comparison). Depending on the desired outcome of the analysis and the size of the dataset, these runtimes might become a limiting factor.

Table 4.2.6. Runtime evaluation of binding site comparison methods with respect to dataset 6. The numbers of successfully prepared pockets and the number of comparisons are given in parentheses, respectively. Exemplary runtimes for methods that demand a pre-processing by the user are highlighted by an asterisk. The average runtimes are coloured by a gradient ranging from green (minimum runtime) to yellow to red (maximum runtime).

Method	Data Basis	Preparation Runtime [s] (Number of Structures)	Comparison Runtime [s] (Number of Comparisons)	Total Runtime [s]	Average Runtime per Comparison [s]
PocketMatch[49]	distance lists	28.97*	0.28	29.25	0.000028
KRIPO[74]	fingerprint	446.50	0.92	447.42	0.000092
RAPMAD[270]	histogram	71.42 (100)	2.36 (8,281)	73.78	0.000285
FuzCav[73]	fingerprint	399.88 (96)	5.59 (9,216)	405.47	0.000607
FuzCav(PDB)[73]	fingerprint	236.73 (96)	5.64 (9,216)	242.37	0.000612
TM-align[71]	matrix	25.72*	65.96	91.68	0.006596
Shaper(PDB)[50]	3D points (grid)	181.16 (96)	364.42 (9,216)	545.58	0.039542
Shaper[50]	3D points (grid)	384.21 (96)	367.21 (9,216)	751.42	0.039845
VolSite/Shaper[50]	3D points (grid)	537.00 (76)	248.77 (5,776)	785.77	0.043070
ProBiS[61]	graph	6.95	479.32	486.27	0.047932
VolSite/Shaper (PDB)[50]	3D points (grid)	259.54 (57)	162.26 (3,249)	421.80	0.049942
TIFP[268]	fingerprint	228.30 (77)	550.88 (5,929)	779.18	0.092913
TIFP(PDB)[268]	fingerprint	194.36 (47)	205.56 (2,209)	399.92	0.093056
Grim(PDB)[268]	graph	169.33 (96)	1,714.49 (9,216)	1,883.82	0.186034
Grim[268]	graph	220.17 (95)	2,104.99 (9,025)	2,325.16	0.233240
IsoMIF[269]	graph	752.83	2,561.44	3,314.27	0.256144
SiteHopper[272]	3D points	154.01	3,828.61	3,982.62	0.382861
Cavbase[64]	graph	67.89 (100)	21,823.71 (8,281)	21,891.60	2.635396
SMAP[69]	graph	1.69	42,346.74	42,348.43	4.234674
SiteEngine[62]	3D points	328.81	81,193.54	81,522.35	8.119354
SiteAlign[77]	fingerprint	28.97*	286,326.41	286,355.38	28.632641

For a comparison of one binding site of interest against all known ligand binding sites as stored in the sc-PDB[10] (9,283 entries) on a single CPU, SiteAlign comparisons would take approximately three days, while PocketMatch will perform these comparisons in 0.28 seconds. RAPMAD is approximately 9,200 times faster than its graph-based counterpart Cavbase. In contrast, TIFP is only two times faster than the graph-based method Grim. Given such small runtime differences, the use of the better performing method is advisable.

The runtime also depends on the type of comparisons. Pairwise comparisons were performed with Grim, IsoMIF, SiteAlign, SMAP, TIFP, TM-align, and VolSite/Shaper, i.e., each tool was invoked separately for each pairwise comparison. This might become the time-limiting step for some methods. Cavbase, FuzCav, PocketMatch, ProBiS, RAPMAD, SiteEngine, and SiteHopper enable the comparison of one query against a list of targets in one run. This causes a speedup as the tool has to be invoked only once.

For ProBiS, the conversion of PDB files to the surface file format is necessary. Thus, the cavity preparation time reflects this preliminary step.

Additionally, KRIPO and RAPMAD enable even faster comparisons by providing SQLite databases of the modelled sites which enable the automated calculation of similarity matrices. These factors illustrate the difficulty in analysing and comparing runtimes of the comparison methods. In this analysis, the fastest option available for each tool was chosen. The computational cost will therefore differ depending on the implementation of the tool and the type of employed approach (comparisons of distinct binding site pairs, query-based analyses, generation of similarity matrices, etc.).

Finally, the composition of the dataset has to be considered. Changing the size of the binding sites might have a significant impact on the time per comparison. While the runtimes of fingerprint-based methods will not be highly influenced by these factors, graph-based methods might become significantly slower with an increasing number of vertices and edges. Cavbase, for example, does not compare very large binding sites.

4.2.8 Conclusion

Plenty of factors influence the decision for a suitable binding site comparison tool. Table 4.2.7 summarises the most important criteria to decide on a tool and provides an estimate of the performance of the analysed methods.

First, the necessary structure pre-processing steps might have an impact on the choice. The preparation of huge datasets, e.g., for the prediction of potential off-targets using all known binding sites in the PDB, requires tools that enable an automated binding site processing and annotation. Therefore, it is important that binding sites are prepared in an automated manner. The KRIPO developers provide scripts to automatically process the binding sites of interest. The same holds true for FuzCav, Grim, TIFP, Shaper, and VolSite/Shaper. Nevertheless, this step is not completely automated as for some PDB structures, the corresponding MOL2 files had to be generated externally. Cavbase and RAPMAD require XML-formatted binding sites as input which can be automatically retrieved with the help of the CSD Python API. Second, the inclusion of all available binding sites might be crucial for analyses such as the prediction of off-targets and the prediction of druggable binding sites. To this end, methods with low failure rates should be chosen.

Sometimes, it is also intended to analyse the similarity of predicted binding sites to already known ones, e.g., for the identification of novel promising targets. Except for IsoMIF, the interaction-based methods cannot be used for the comparison of predicted cavities. SiteEngine will also fail as the tool relies on the construction of protein- and ligand-surfaces to compare the resulting site surfaces. Furthermore, SMAP does not process detected apo sites. The generation of putative ligand binding modes for known ligands via docking or pharmacophore searches might provide suitable structures for these methods. Relating back to Section 4.1.2, some energy- and knowledge-based pocket prediction tools (e.g., G-LoSA, COACH, Binding Response, or MolSite) report ligands in the identified sites.

For Cavbase, FuzCav, RAPMAD, Shaper, and IsoMIF, binding site identification tools are provided together with the comparison software. PocketMatch, ProBiS, SiteAlign, and TM-align process user-defined binding sites based on the underlying residues. Therefore, it is possible to apply them to externally predicted binding sites. SiteHopper was also shown to process predicted pockets if fpocket[196]-derived cavities were utilised as pseudo-ligands.[372]

Table 4.2.7. Quality criteria for binding site comparison methods. A plus in parentheses means that the predicted binding site has to be given with the corresponding coordinates of a binding site (BS) prediction as "artificial" ligand, a simple plus denotes tools that offer a way of binding site prediction. With respect to the runtime evaluation, "+", "/", "-" denote comparison algorithms that require several ns, µs, or s per comparison, respectively. With respect to the scoring, a "+" was assigned to those tools where the intervals of upper and lower whiskers of active and inactive pairs did not overlap. A "/" denotes tools where the upper and lower quartile for the active and inactive pairs do not overlap. The plus sign in brackets for SiteAlign with respect to the ROCS dataset indicates the score dependence of this rating. With respect to other factors, tools that were clearly outperformed by many other tools were assigned a "-".

Method	BS Preparation (Ease)	BS Preparation (Completeness)	Applicability towards Predicted BS	Computational Cost	BS Definition (Differing Ligands)	BS Definition (Ranking)	BS Flexibility	BS Properties (Ranking)	ROCS Dataset	Dataset of Successful Applications	Visualisation
Cavbase	+	-	+	-	+	+	+	+	-	+	+
FuzCav	/	+	+	+	+	/	+	+	-	+	-
Grim	/	-	-	/	-	-	+	-	/	-	+
IsoMIF	+	+	+	/	-	-	-	-	+	+	+
KRIPO	+	+	-	+	-	/	+	+	/	+	+
PocketMatch	-	-	(+)	+	-	/	+	-	+	+	-
ProBiS	+	+	(+)	+	+	+	+	-	-	+	+
RAPMAD	+	-	+	+	-	-	-	+	-	-	-
VolSite/Shaper	/	-	+	/	+	/	-	+	-	+	+
SiteAlign	-	+	(+)	-	+	+	+	+	(+)	+	+
SiteEngine	+	+	-	-	+	/	+	+	-	+	+
SiteHopper	+	/	(+)	/	+	+	+	+	+	+	+
SMAP	+	+	(+)	-	+	+	+	+	-	+	+
TIFP	/	-	-	/	-	-	+	-	-	-	-
TM-align	-	+	(+)	/	+	+	+	n/a	+	+	+

The computational cost is valuable information when it comes to the analysis of large datasets as necessary for the prediction of off-targets or the analysis of site relationships by comparing structures in an all-against-all manner. However, the drawbacks of extremely fast methods must be considered.

Based on the analysed datasets, we can also draw conclusions with respect to a suitable input for binding site comparison. Tools with an inferior performance for the dataset of structures with identical sequences (dataset 1) should not be used if only one or a few structures of protein-ligand complexes are known. This holds especially true for interaction-based algorithms.

Possibly, the problem can be circumvented by using docking-derived binding site structures. With respect to the sensitivity towards protein flexibility, it is advisable to use all available crystal structures or multiple NMR models to retrieve promising results (dataset 2). Although the quality of NMR structures is difficult to assess,[409] they served as valid structural models for some structural biology and computational modelling projects.[410,411] The G-factors in Table 4.2.2 indicate inaccuracies of the NMR ensemble structures of dataset 2, but the mean G-factor hints at predominantly usual dihedral angles and main-chain covalent forces in the models. Additionally, the generation of MD simulation-derived representative structures which can be used as input for the comparisons might be beneficial. In the case of restricted structural knowledge, one of the methods with superior performance on datasets 1 and 2 should be used.

The scoring schemes of some methods suffer from a small score difference window between the score distributions for similar and dissimilar binding sites. The number of tools which can distinguish between similar pockets with minor dissimilarities is small. Moreover, only a subset of the tools (SiteAlign, SMAP, SiteEngine, SiteHopper, and KRIPO) was able to sufficiently penalise differences between protein structures and their artificially designed variants. Therefore, hierarchical comparison schemes taking advantage of different methods might provide improved results. However, tools which are not able to detect minor binding site differences might be successfully applied in the elucidation of similarities between the distant, but related cavities of proteins with unknown similarities.

The dataset of ROCS structures revealed that especially residue-and surface-based tools do not identify challenging, solely interaction-based binding site similarities. It is obscure whether they might perform better when using various structures to consider binding site definition and flexibility. Even minor side-chain movements affect the feature placement. However, the outcome of this analysis is not universally applicable as the question of how to reliably score binding site pairs without complex interaction patterns and physicochemical characteristics persists.

The dataset of successful applications gives some insight into the basic applicability of the tools for different scenarios. All tools showed a high early enrichment. The analysis of the performance for this dataset suggests that the individual strengths and weaknesses of all tools finally level out to overall promising performance. Finally, the evaluation of the tools with regard to the dataset of Barelier *et al.*[267] shows that in most cases the visual inspection is inevitable if the user is interested in a drug repurposing strategy or wants to predict polypharmacology. Therefore, a plus or minus sign is assigned to methods depending on the possibility to visually analyse the resulting binding site alignments.

Although some tools have never been applied in independent research projects, the results presented here show that all approaches have unique strengths that can be utilised in numerous projects. Several studies including Grim,[412] RAPMAD,[413] and SiteHopper[414] underline their unique capabilities. However, the tools Grim and TIFP show comparatively low applicability to the identification of binding site similarities.

With respect to both tools, it must be stated that their original application domain was the rescoring of docking poses, e.g., in an SBVS campaign or for pose selection. For these approaches, both tools perform reasonably well, as shown in the original publication[268] and in the D3R grand challenge 2015.[412] Additionally, the original structures as stored in the sc-PDB[10] could not be used as the datasets of this benchmark study include sc-PDB- and PDB-derived structures. Restricting the binding site comparisons with TIFP and Grim to solely to the structures derived from the sc-PDB might provide another perspective on the performance of both tools as their results highly depend on the complete pre-processing of the binding sites.

It could also be shown that the application of score cut-off values to discriminate between similar and dissimilar binding sites is questionable. Although it is sometimes the only possibility to detect related sites, the assignment of appropriate thresholds is challenging and an elaborate cut-off adjustment with respect to the specific study is crucial.[53] The investigations herein highlight the strong context dependence of such cut-offs (see Table A2.31). If the classification of similar and dissimilar pairs is necessary, the appropriate cut-off has to be evaluated with a dataset that reflects the main purpose of the analysis. Score thresholds for the decoy datasets (datasets 3 and 4) should be considered if binding sites with minor dissimilarities are compared. In contrast, the elucidation of local binding site similarities or similarities of pockets of unrelated proteins benefits from the less restrictive cut-offs that were determined for datasets 6 or the dataset of lifelike pocket similarities (dataset 8). The methods SMAP and ProBiS offer additional statistical measures to estimate the significance of a binding site match. Nevertheless, it has to be emphasised that the application of thresholds often means disregarding potentially interesting and yet unexplored binding site similarities.

Some guidance can be provided regarding several aspects of binding site comparison, even though it is not possible to identify a single method which satisfies all needs. The impact of the ease and completeness of binding site pre-processing steps increases with a growing number of sites for comparison. While both factors are crucial for elaborate projects, e.g., off-target prediction or the identification of novel binding sites, their importance decreases for comparatively small datasets used for selectivity profiling or polypharmacology analyses. An application to predicted binding sites is relevant for off-target predictions, drug repurposing projects, or the identification of novel targets. For the establishment of polypharmacology, one generally refers to known druggable binding sites. The analysis of similar binding sites to elucidate minor structural differences is highly influenced by the binding site definition and flexibility, as is the ranking with respect to binding site properties. In contrast, the discovery of similarities between the cavities of unrelated proteins does not necessarily depend on an accurate scoring. MD simulations or the use of NMR ensembles might help to circumvent potential failures of the binding site comparisons due to insufficient consideration of protein flexibility. If it is necessary to disclose the similarity between functionally unrelated binding sites that bind to similar ligands for understanding polypharmacology or off-target effects, a visualisation of the identified similarities is crucial. In contrast, it is of minor importance for selectivity profiling or the analysis of evolutionary relationships.

These criteria should become the focus of future benchmark analyses of further promising site comparison approaches. Such evaluation studies are the only means to guide the rational choice of a method.

Combined approaches might turn out most beneficial for different projects. As seen for the dataset of successful applications, most methods were suitable to enrich similar site pairs in lists that are ranked by the similarity scores of the pairs. However, we observe complementary results for the different approaches. While interaction-based methods are especially reliable for the detection of similar binding site pairs of unrelated proteins, residue- and surface-based methods offer more promising scoring schemes and are less dependent on the binding site-occupying ligands. Figure 4.2.21 summarises several types of binding site comparison approaches with the corresponding field of research they were successfully applied in. Maybe a consensus approach, incorporating all types of binding site modelling approaches, will pave the way to even more reliable binding site comparisons.

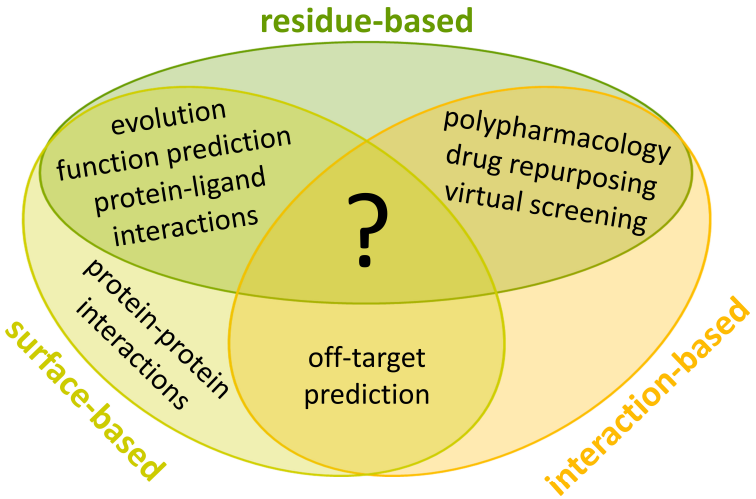


Figure 4.2.21. Binding site comparison applications in relation to different binding site modelling approaches. The Venn diagram of successful applications of the benchmarked residue-, surface-, and interaction-based tools in different research scenarios is given. The diagram was generated using DrawVenn.[415]

4.3 Secondary Structure Assignment

4.3.1 Motivation

The assignment of helices and β-sheets in protein structures can be achieved by numerous methods (see Section 1.2.1). Some of these methods solely rely on the backbone geometry of the involved residues (e.g., ASSP[122] and SEGNO[119]), while other tools consider the backbone hydrogen bonds within the respective SSEs (e.g., DSSP[104] and STRIDE[105]). SHAFT (Secbase automated Helix Assignment adapted From Turns)[106] is a method that assigns helices based on hydrogen-bonded normal and open turns that are characterised by distinct dihedral angles. The method assigns right-handed α-, 3_{10}-, π-, and the so-called γ-helices. The missing β-strand assignment and further flaws of the method (Section 4.3.3) encouraged us to develop a novel comprehensive SSE assignment method based on turn structures that are commonly found in protein structures.

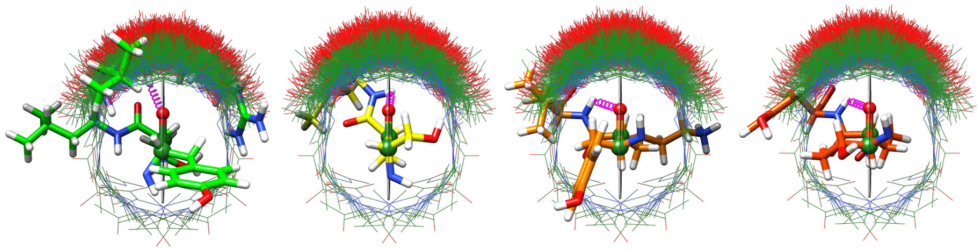
4.3.2 Turn Classification

Assignment of Hydrogen-Bonded Turns

The classification of normal and reverse turn types in proteins requires a criterion for the detection of hydrogen bonds, the calculation of their energies, and an energy cut-off to exclude weak hydrogen bond interactions. In Secbase, this was achieved by using the hydrogen bond definition by Kabsch and Sander which was applied for the SSE classification with DSSP.[104] The occurrence of normal 3 turns within β-sheets gave rise to a critical reconsideration of the calculation of the underlying hydrogen bond potentials. Turns of the category normal 3 (hydrogen bond between the residues at the positions *i* and *i+2*) are found in a large number of β-strands as assigned by DSSP. This turn type is used for the assignment of γ-helices by SHAFT leading to overlapping strand and helix definitions. A closer examination of the corresponding turns revealed some inadequacies of the applied potential as illustrated in Figure 4.3.1. The four normal 3 turns that were classified with Secbase[416] show highly different geometries. The search for spatial characteristics of hydrogen bond interactions between carbonyl oxygen and amide nitrogen atoms in the knowledge-based interaction database IsoStar[417] shows a distinct distribution of angles and distances. Three of the four exemplary turns do not seem to be in accordance with the IsoStar[417] analysis-based scatterplots. Nevertheless, the assigned energies are high compared to that of a turn that fits better to the scatterplot of known hydrogen bond interactions observed in the PDB and CSD (Figure 4.3.1, green turn structure).

An exhaustive evaluation of hydrogen potentials was infeasible and clearly exceeds the goal of reliably assigning turn structures in proteins. Therefore, we restricted the search for a more appropriate hydrogen bond assignment to five empirically derived hydrogen bond potentials which were used to analyse and assign SSEs or to improve the quality of protein crystal structures. The focus was drawn to a clear distinction between geometrically unfavourable and favourable hydrogen bonds. A visual inspection of assigned hydrogen bonds for different turn structures further substantiated the final decision for the definition as published by Dahiyat and co-workers.[286]

The corresponding energies for the four exemplary turn structures are given in Figure 4.3.1. While the definition by Kabsch and Sander and a related approach by Zacharias[281] are not suitable to penalise geometrically unfavourable hydrogen bond interaction, the remaining approaches show a clear preference for the normal 3 turn with an optimum geometry. Nevertheless, the clearest distinction was achieved using the definition applied by Dahiyat and colleagues. In contrast to the original DREIDING potential, the additionally set cut-off values enable a reliable assignment of hydrogen-bonded turns. Moreover, the potential already led to convincing results in the past and is frequently applied in protein design.[418–421]



	E_{hb} [kcal/mol]	E_{hb} [kcal/mol]	E_{hb} [kcal/mol]	E_{hb} [kcal/mol]
Kabsch & Sander, 1983	-2.54	-0.06	-0.73	-0.47
Zacharias, 2015	-3.72	0.23	-1.27	-0.97
Boobbyer *et al.*, 1989	-2.48	-2.62	-1.33	-0.81
Mayo *et al.*, 1990	-6.13	-0.02	-0.01	0.00
Dahiyat *et al.*, 1997	**-4.69**	**-0.04**	**0.00**	**0.00**
Fabiola *et al.*, 2002	-0.58	-0.25	0.00	0.00

Figure 4.3.1. Hydrogen bond energies and geometry. Different normal 3 turns (assigned based on the DSSP hydrogen bond definition) with varying geometric properties (stick representation) and the IsoStar[417]-derived geometric distributions of hydrogen bonds between carbonyl oxygen and amide nitrogen atoms (wire representation) are presented. The green normal 3 turn is an example of an expected structure, while the yellow, orange, and red turn structures represent unanticipated hydrogen bond assignments (pink springs). Additionally, the corresponding hydrogen bond energies (E_{hb}) derived from the analysed hydrogen bond potentials are given.

The original E_{hb} threshold of -2 kcal/mol for a hydrogen bond to exist was adjusted to -0.5 kcal/mol. This adjustment is based on an analysis of hydrogen bond geometry in combination with the consistency of the subsequent secondary structure assignment. As previously shown, an increase in the E_{hb} threshold leads to an increased consistency with respect to secondary structure classification.[111] Minor geometric irregularities are tolerated in favour of a reliable helix and sheet assignment. Moreover, the visual inspection of turn structures with a hydrogen bond potential between -0.5 and -2 kcal/mol confirmed the validity of the final decision. Together with the optimised angle and distance cut-offs, this hydrogen bond potential ensures the retention of crucial turn structures, while structures such as normal 3 turns with N-H-O angles of below 130° (as part of γ-helices) are neglected.

Assignment of Open Turns

The original Secbase[416] implementation assigns open turns for Cα-Cα distances between 4 and 10 Å. To reduce the size of the dataset for turn classification and, nevertheless, assign reliable helix termini using the novel turn classification, we analysed the impact of open turns with distinct Cα-Cα distance ranges on the classification of helices for the structures in the training set.

For the capping of 3_{10}-helices, open 4 I turn structures are considered in the SHAFT-based helix assignments. The distances of all open 4 I turns were binned. In the case of α-helices, we repeated this analysis for the open 4 I and open 5 II turns. Figure 4.3.2 presents the result of this analysis. In summary, the Cα-Cα distances for most open turns that are involved in the helix assignments range between 5 and 8 Å. Nonetheless, open 4 I turns with larger distances are found. A visual inspection of the corresponding helices revealed highly irregular helix caps. Therefore, we decided to define open turns in a distance range between 4 and 8 Å.

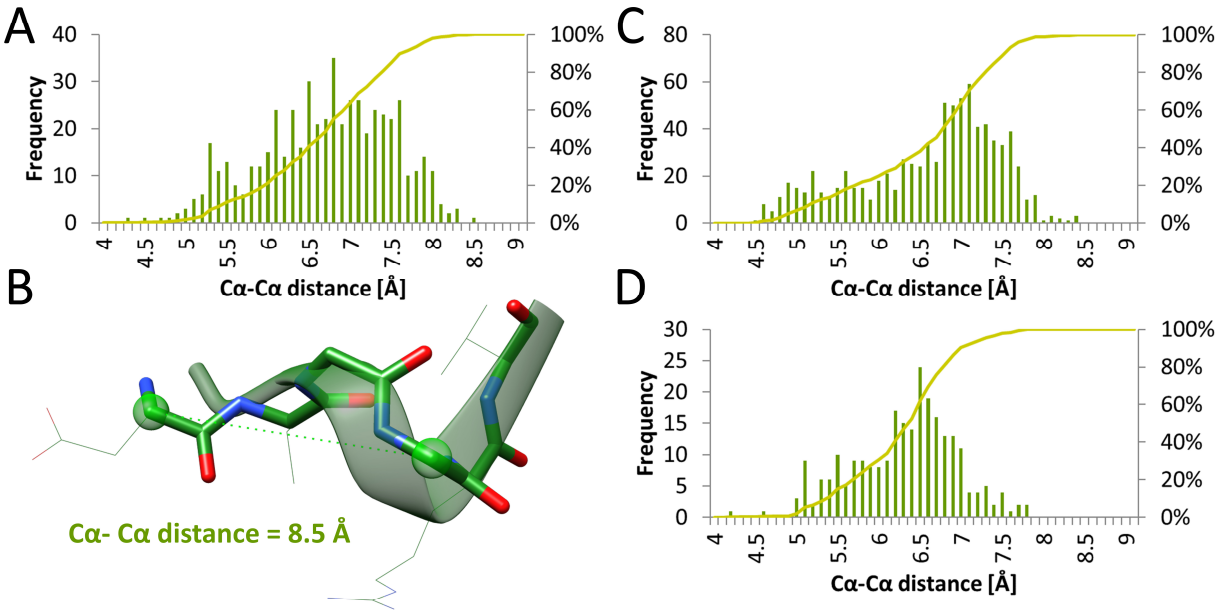


Figure 4.3.2. Cα-Cα distances of open turns used for helix capping in SHAFT. (A) The binned distances for the open 4 I turns that are used for the SHAFT-based capping of 3_{10}-helices are shown. (B) The N-terminal cap of a 3_{10}-helix (residues 223 to 228 in the protein structure with the PDB-ID 3mcz, chain A) is shown. The green spheres highlight the terminal Cα atoms of the capping open 4 I turn whose distance exceeds 8 Å. The binned distances for the open 4 I (C) and open 5 II (D) turns that are used for the α-helix capping are presented. In panels A, C, and D the binned frequencies are given as dark green bars, while the cumulative frequency is given by the light green lines.

Dataset Generation

Due to the modifications outlined above, the number of assigned turns per structure will be smaller than that for Secbase. Moreover, the size and composition of the dataset have a huge impact on the final outcome of the analysis of amino acid propensities.[422] Therefore, we generated a sequence-clustered dataset of high-quality protein structures by applying the sequence-culling server PISCES.[258] The identity threshold for culling was set to 35% to ensure the generation of a sufficiently large dataset (3,597 single chain structures) to ensure the coverage of all turn categories. Accordingly, a test set was generated to probe the external applicability of the developed ESOM-based classification.

The SCOT-based classification differentiates between turn families (hydrogen bond- or distance-based turns), turn classes (turn lengths), and turn types (turns with distinct conformations within the respective category). The following discussion will first focus on turn categories which play a significant role in the helix assignment with SCOT. Finally, reverse turn structures and their impact on the assignment of SSEs in proteins will be discussed.

ESOM-Based Turn Clustering

The following sections are partially dedicated to a comparison between the novel turn classification and the former turn classification with Secbase. The basic differences are the preparation of the input structures, the assignment of the hydrogen bonds, the angle transformations applied for the ESOM-based clustering, and the significantly increased sizes of the trained ESOMs. These modifications lead to a refined and more reliable classification of turn types in the dihedral angle space and the support of modified residues and D-amino acids. Some more specific differences between the type assignments will be discussed in the following sections. To this end, turn types assigned by the old classification with Secbase and the novel assignment are projected on the new ESOMs to enable an immediate comparison.

The new classification of turns based on ESOM-derived U-Matrices led to reasonable turn types for the hydrogen-bonded as well as for the open turns. A comparison to the formerly defined turn types with Secbase reveals that the increased size of the training set and the increased multiplier applied for ESOM training led to a higher number of distinct turn types. This is most prominent for open turn structures which are sterically less restricted than the hydrogen-bonded ones. Consequently, only the highest occupied turn types are discussed in the following sections, while Tables A3.1-19 in Appendix 3 include all data for all assigned turn types (the dihedral angles, hydrogen bond energies E_{hb}, Cα-Cα distances, and residue preferences for the assigned turn types in the training set).

Turn Classes and their Meaning for the Assignment of Secondary Structure Elements

The following paragraphs discuss the classification of turns also with respect to their use in the developed SSE assignment method SCOT. For convenience, we use the number of residues for the turn nomenclature and avoid the use of the original names ε- (2 residues), γ- (3 residues), β- (4 residues), α- (5 residues), and π- (6 residues) turns. The turn type numbers are assigned based on the cluster sizes. The turn types will be discussed with a special focus on normal and open turn types which are mainly used for the helix assignments with SCOT.

Normal Turns

Normal 3 Turns

The normal 3 turn category is represented by a small number of turn structures. Only two types are distinguished: n3-1 (formerly inverse γ-turn) and n3-2 (formerly normal γ-turn). The dihedral angles of turn type n3-1, which is more frequently observed than the normal γ-turn, correspond to those of the left-handed 2.2_7-helix.[84,423] In contrast, the n3-2 turn type can be used to assign right-handed 2.2_7-helices. The corresponding U-Matrix of the ESOM training with the projected types of the new and the former assignment shows that the assignment of this study nearly perfectly matches the previously established one (Figure 4.3.3). Nevertheless, it has to be stated that the initial training for the Secbase-derived classifications was conducted after assigning hydrogen bonds with the criterion defined by Kabsch and Sander.[104] This led to a high number of turns for this category as compared to the current study. Due to their low hydrogen bond energies, as discussed in the first paragraphs of this section, most Secbase-assigned normal turns were not classified with SCOT.

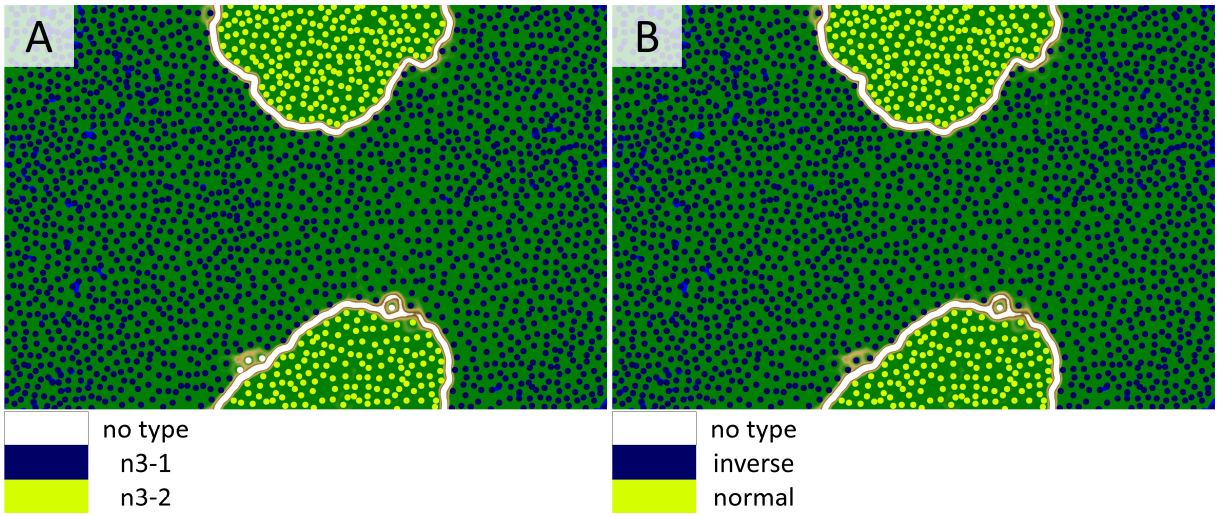


Figure 4.3.3. Normal 3 turn clustering. The U-Matrix for normal 3 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown.

Table 4.3.1 gives the angular means and standard deviations as well as the mean hydrogen bond energy of the two normal 3 turn types assigned in this study. The low hydrogen bond energy for both types compared to normal turns with four to six residues as well as the residue preferences for Pro (n3-1) and Gly (n3-2) residues (see Table A3.1) explain why normal 3 turns are rarely observed in protein structures.

For the highest occupied type n3-1, a high prevalence of aromatic and lipophilic residues at positions *i* and *i+2* and a high frequency of Pro residues at position *i+1* is found. Both types are characterised by frequently occurring Ser and Thr residues at position *i+2*. In contrast, high frequencies of Gly residues at positions *i* and *i+1* and of aromatic residues at position *i+1* are only observed for type n3-2 (see Table A3.1).

Table 4.3.1. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of normal 3 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}		E_{hb} [kcal/mol]	
		∅	±	∅	±	∅	±	∅	±	∅	±
n3-1	1,852	-175.8	4.8	-82.3	5.7	61.1	9.7	178.3	4.7	-1.0	0.4
n3-1	*1,273*	*-175.6*	*4.9*	*-82.4*	*5.4*	*61.4*	*9.1*	*178.3*	*4.8*	*-1.0*	*0.4*
n3-2	334	175.6	5.7	74.3	6.5	-52.0	10.9	-178.1	5.4	-1.4	0.7
n3-2	*228*	*175.8*	*5.8*	*73.8*	*6.1*	*-51.0*	*10.9*	*-177.7*	*5.3*	*-1.4*	*0.7*

Normal 4 Turns

Seven normal 4 turn types were assigned based on the underlying U-Matrix (Figure 4.3.4). The following correspondences to the formerly assigned turn types were identified: n4-1 and I, n4-2 and II, n4-3 and I', n4-4 and II', n4-5 and VIa1, and n4-6 and VIb. The novel type n4-7 is represented by a small cluster in the U-Matrix which is situated in between the types n4-1 and n4-2 but shows distinct dihedral angle ranges (Table 4.3.2). Broken turns were not identified in our training set.

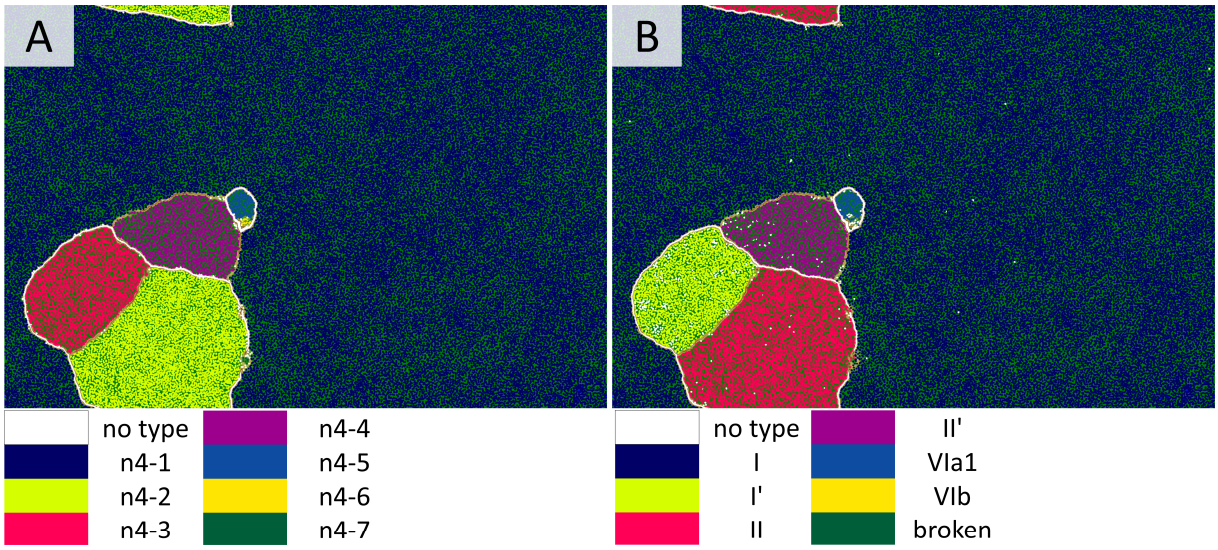


Figure 4.3.4. Normal 4 turn clustering. The U-Matrix for normal 4 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown.

A *cis* peptide bond between the residues at the positions *i+1* and *i+2* is observed in the normal 4 turn types n4-5 and n4-6 which is in line with their correspondence to the original types VIa1 and VIb. Both turn types were assigned as one type by the Secbase classification (a1 and b). However, the dihedral angle ranges of these types suggest major conformational differences and the distribution of over- and underrepresented residues is significantly different for both turn types (see Table A3.2 in Appendix 3).

A closer examination of Table 4.3.2 and the Ramachandran plots presented in Figure 4.3.5 reveals that the turn types n4-1 and n4-3 show dihedral angle ranges that are typical for right-handed (φ = -62.8°, ψ = -16.5°)[424] and the corresponding left-handed (φ = +62.8°, ψ = +16.5°) 3_{10}-helices. Gly residues are overrepresented in the latter turn type, together with other positive and negative ionisable residue types (see Table A3.2). Intriguingly, the hydrogen bond energies for both turn types are less stable than the ones for other normal 4 turns. Polar side-chains of residues in n4-1 and n4-3 turns might additionally stabilise their conformation.

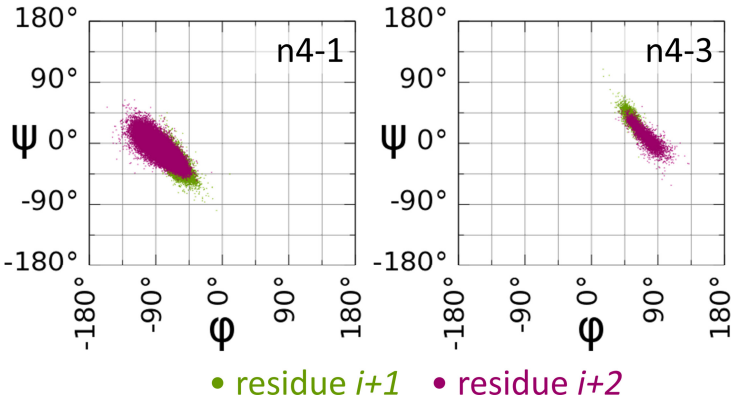


Figure 4.3.5. Ramachandran plots for the normal 4 turn types used for the 3_{10}-helix assignment. Turn type n4-1 (left panel) is found as a repetitive motif in right-handed 3_{10}-helices, while successive turn structures of type n4-3 (right panel) are characteristic of left-handed 3_{10}-helices.

Table 4.3.2. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of normal 4 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}		φ_{i+2}		ψ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
n4-1	70,498	-178.7	4.4	-61.6	6.8	-28.4	11.4	179.7	3.7	-78.2	15.9	-11.9	15.2
n4-1	*48,974*	*-178.7*	*4.3*	*-61.6*	*6.9*	*-28.5*	*11.4*	*179.7*	*3.7*	*-78.0*	*15.9*	*-12.1*	*15.2*
n4-2	6,970	179.4	5.1	-56.5	6.5	131.9	7.4	179.3	3.5	81.8	13.7	-0.3	14.8
n4-2	*4,854*	*179.4*	*5.0*	*-56.5*	*6.6*	*131.9*	*7.7*	*179.4*	*3.5*	*81.7*	*13.8*	*-0.3*	*14.8*
n4-3	3,227	179.3	5.0	53.9	6.6	37.1	10.8	178.2	3.7	75.3	12.9	8.6	14.9
n4-3	*2,248*	*179.4*	*5.4*	*53.9*	*6.7*	*37.1*	*10.9*	*178.2*	*3.6*	*75.2*	*12.8*	*8.7*	*14.8*
n4-4	2,166	-179.1	6.4	56.8	7.9	-126.3	9.3	-179.3	3.5	-85.1	17.4	-0.1	17.1
n4-4	*1,533*	*-179.1*	*6.1*	*56.8*	*7.8*	*-125.9*	*9.5*	*-179.5*	*3.6*	*-84.6*	*17.5*	*-1.0*	*17.6*
n4-5	212	176.9	6.3	-54.2	8.0	141.0	5.9	5.6	5.2	-91.0	7.6	16.4	10.9
n4-5	*158*	*177.3*	*6.0*	*-54.4*	*8.2*	*140.9*	*5.7*	*5.1*	*4.6*	*-91.1*	*8.1*	*16.7*	*10.6*
n4-6	17	-179.7	6.1	-87.8	9.7	130.3	10.9	-4.0	7.6	-116.3	13.8	68.4	13.0
n4-6	*15*	*179.5*	*8.7*	*-88.9*	*13.7*	*131.5*	*11.6*	*-6.0*	*13.3*	*-110.9*	*15.1*	*67.9*	*20.2*
n4-7	9	-177.5	6.5	-72.5	9.9	76.8	9.6	-175.1	7.6	179.3	17.0	-32.5	10.4
n4-7	*4*	*171.7*	*1.3*	*-67.0*	*8.8*	*85.4*	*8.1*	*-172.0*	*8.0*	*169.7*	*4.8*	*-32.1*	*10.5*

type	|turns|	ω_{i+2}		E_{hb} [kcal/mol]	
		∅	±	∅	±
n4-1	70,498	-179.7	4.9	-1.7	0.7
n4-1	*48,974*	*-179.7*	*4.8*	*-1.7*	*0.7*
n4-2	6,970	179.4	4.9	-3.0	1.2
n4-2	*4,854*	*179.3*	*4.8*	*-3.0*	*1.2*
n4-3	3,227	-179.9	4.6	-1.9	0.7
n4-3	*2,248*	*-180.0*	*4.5*	*-1.9*	*0.7*
n4-4	2,166	-179.6	5.4	-3.1	1.2
n4-4	*1,533*	*-179.6*	*5.4*	*-3.1*	*1.2*
n4-5	212	179.7	6.1	-2.8	1.0
n4-5	*158*	*-179.9*	*5.7*	*-2.8*	*1.1*
n4-6	17	-173.0	7.2	-3.7	1.6
n4-6	*15*	*-180.0*	*13.0*	*-3.4*	*1.6*
n4-7	9	176.0	4.3	-2.9	1.4
n4-7	*4*	*179.9*	*10.5*	*-3.0*	*1.2*

Normal 5 Turns

The resulting U-Matrix for turns of the ESOM training based on the dihedral angles of the normal 5 turn category (Figure 4.3.6) shows a close correspondence between the initially defined turn types and the new ones. The highest occupied type n5-1 corresponds to type I. Types n5-2 and n5-7 were formerly defined as type II by Secbase. Most turns of type n5-3 were classified as type VI, which is characterised by a *cis* peptide bond between the residues at the positions *i+1* and *i+2*. The original turn types IIIa and IIIb are now part of one unique turn type (n5-4) but are additionally in the clusters of types n5-11 and n5-12. Further correspondences are: n5-5 and n5-14 to type IV, n5-6 to type III', n5-8 and n5-10 to type V, and n5-9 to type I'. Turns of type n5-14 remained unclassified (no type) in the turn classification by Secbase.

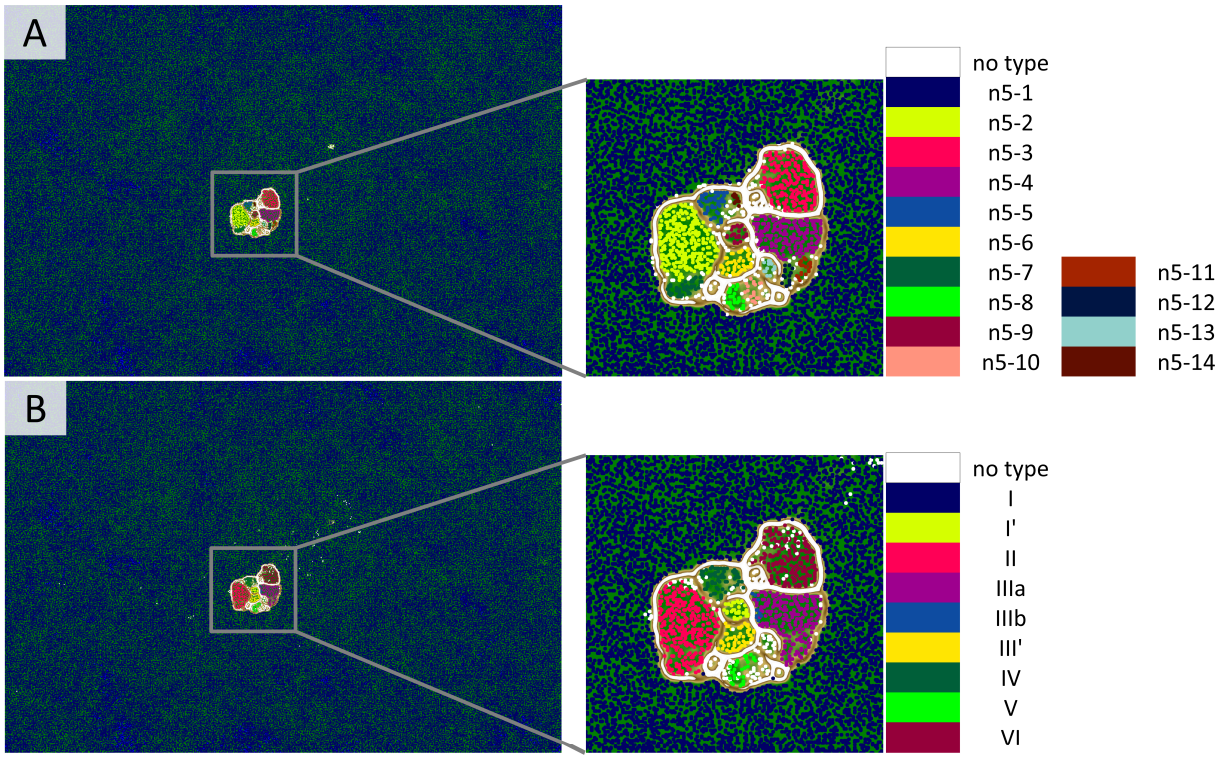


Figure 4.3.6. Normal 5 turn clustering. The U-Matrix for normal 5 turns and with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown.

The dihedral angles of the different classified turn types show distinct ranges and low standard deviations (Table 4.3.3). Turns of the former types II, III, IV, and V, which were characterised by high standard deviations of the dihedral angles, could be subdivided with the help of the novel U-Matrix which was derived from an ESOM training with a multiplier of 5 (0.5 in the training for the turn assignment with Secbase). This leads to the overall lower standard deviations and the more robust classification of normal 5 turn types compared to Secbase.

Higher hydrogen bond energies compared to the normal 3 and 4 turn types are observed which can be attributed to the higher number of conformational degrees of freedom and the occurrence of further stabilising hydrogen bonds (see Reverse Turns).

The turn types with the highest hydrogen bond energies are n5-1, n5-13, and n5-14. Gly residues are overrepresented in turn types n5-2, n5-4, n5-5, and n5-6 (see Table A3.3), which partially explains the comparatively high standard deviations compared to other normal 5 turn types. The same holds true for type n5-3 which is frequently occupied by Pro residues. It is the turn type with the highest standard deviations of the dihedral angles. The clustering of this type to further sub-types might be possible. However, we abandoned this idea to avoid a too fine-grained discrimination. Further discriminations might be possible by using a larger dataset with more neurons to the disadvantage of a higher computational cost.

The average dihedral angles of the normal 5 turn types shed light on the turn conformations whose angles correspond to the dihedral angle ranges of α-helices (Table 4.3.3 and Figure 4.3.7). The consecutive dihedral angles of the turn structures of type n5-1 (formerly type I) are in the range of those of right-handed α-helices (φ = -57° and ψ = -47°).

The turn structures of type n5-9 (formerly I'), which are scarce, are characterised by dihedral angles which correspond to those of left-handed α-helices (φ = +57°, ψ = +47°). The latter turn type is characterised by a lower mean hydrogen bond energy, which is in accordance with the sparseness of left-handed helices in proteins composed of L-amino acids due to their conformational restrictions. In consequence, the use of the n5-1 and n5-9 turn types for the assignment of α-helices is possible considering their mean dihedral angles.

The rarely detected right-handed ω-helices, which are characterised by dihedral angles of φ = -2.8° and ψ = -54.9°[425] and closely related to α-helices, cannot be identified based on this classification. A finer-grained differentiation of the turns of the n5-1 type cluster might enable their assignment.

Table 4.3.3. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of normal 5 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}		φ_{i+2}		ψ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
n5-1	221,356	179.3	3.5	-62.6	5.4	-41.5	6.7	178.9	3.0	-64.1	6.6	-41.8	6.6
n5-1	*154,102*	*179.3*	*3.5*	*-62.6*	*5.4*	*-41.5*	*6.7*	*178.9*	*3.0*	*-64.1*	*6.6*	*-41.8*	*6.7*
n5-2	283	-178.5	4.7	-57.5	7.4	130.1	6.5	-179.6	3.1	80.1	12.5	-3.2	14.8
n5-2	*208*	*-178.6*	*4.5*	*-57.2*	*6.9*	*130.7*	*6.3*	*-179.5*	*3.2*	*80.0*	*11.6*	*-4.5*	*15.2*
n5-3	219	176.6	6.6	-83.8	21.0	143.5	11.3	2.8	4.7	-90.8	8.3	-2.6	11.4
n5-3	*153*	*176.6*	*6.4*	*-84.1*	*21.3*	*142.7*	*11.9*	*2.9*	*5.0*	*-90.9*	*8.3*	*-2.6*	*12.2*
n5-4	190	175.0	6.3	58.5	9.6	-150.5	16.4	-178.2	3.9	-63.0	6.6	-36.6	7.7
n5-4	*138*	*175.2*	*6.4*	*58.9*	*9.7*	*-151.2*	*15.4*	*-177.4*	*3.7*	*-64.2*	*6.0*	*-36.5*	*7.7*
n5-5	57	178.8	4.7	55.0	6.3	26.9	7.3	-178.6	3.4	63.2	7.3	16.0	9.9
n5-5	*33*	*177.1*	*18.5*	*56.6*	*6.5*	*27.4*	*7.8*	*-177.2*	*3.9*	*63.4*	*10.4*	*14.4*	*12.8*
n5-6	52	-179.5	5.7	-57.5	9.0	150.2	13.3	-178.7	4.6	72.3	15.4	28.0	14.8
n5-6	*37*	*-179.5*	*5.7*	*-57.4*	*7.3*	*150.8*	*11.4*	*-179.4*	*4.4*	*68.3*	*13.1*	*30.4*	*15.3*
n5-7	43	178.2	5.1	-59.4	7.1	142.5	7.3	-179.6	4.7	78.0	9.7	-56.5	8.5
n5-7	*25*	*176.5*	*6.9*	*-57.5*	*8.4*	*141.1*	*8.4*	*178.3*	*7.6*	*78.0*	*9.4*	*-54.9*	*7.5*
n5-8	28	-177.6	4.6	-64.3	8.8	-17.3	11.2	176.9	4.1	-82.4	11.6	-1.2	8.3
n5-8	*19*	*-177.3*	*4.0*	*-64.0*	*7.3*	*-16.2*	*8.6*	*176.5*	*3.2*	*-85.6*	*11.2*	*1.7*	*6.2*
n5-9	31	179.4	6.1	55.4	6.6	44.7	9.3	-179.6	4.0	62.3	8.2	43.9	11.7
n5-9	*27*	*178.8*	*5.8*	*55.4*	*6.1*	*44.1*	*8.2*	*-179.1*	*3.5*	*59.7*	*6.9*	*48.0*	*18.7*
n5-10	20	-178.9	4.2	-66.6	5.2	-22.8	8.9	-179.9	3.1	-98.5	13.3	72.7	15.6
n5-10	*14*	*-178.2*	*3.3*	*-68.2*	*5.6*	*-21.1*	*9.2*	*179.9*	*1.6*	*-97.9*	*12.5*	*67.4*	*11.4*
n5-11	21	-177.6	4.0	52.9	6.4	-137.8	5.7	175.7	4.0	-105.8	11.0	10.6	9.3
n5-11	*11*	*-177.9*	*6.6*	*51.3*	*4.5*	*-133.6*	*6.8*	*176.4*	*4.3*	*-107.4*	*16.7*	*9.6*	*8.8*
n5-12	17	-178.4	4.0	60.9	6.2	-151.2	7.7	-177.1	2.8	-80.3	9.8	-28.3	11.4
n5-12	*9*	*-178.8*	*2.3*	*64.2*	*5.7*	*-153.5*	*9.2*	*-178.0*	*1.5*	*-84.0*	*11.4*	*-25.6*	*9.9*
n5-13	17	177.3	3.7	63.2	8.9	-134.3	10.5	179.2	2.5	-97.2	10.4	74.1	14.7
n5-13	*14*	*177.3*	*4.0*	*61.3*	*7.7*	*-135.2*	*11.2*	*179.1*	*3.7*	*-98.0*	*11.4*	*77.5*	*11.7*
n5-14	9	178.4	5.2	54.3	3.6	35.7	8.5	174.7	4.8	83.6	8.0	-38.3	8.9
n5-14	*3*	*177.0*	*7.5*	*54.1*	*6.3*	*35.4*	*3.5*	*177.1*	*5.8*	*79.3*	*7.9*	*-47.7*	*7.7*

Table 4.3.3 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of normal 5 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+2}		φ_{i+3}		ψ_{i+3}		ω_{i+3}		E_{hb} [kcal/mol]	
		∅	±	∅	±	∅	∅	∅	±	∅	±
n5-1	221,356	179.2	3.4	-65.5	10.4	-40.7	7.7	179.5	3.6	-4.2	1.4
n5-1	*154,102*	*179.2*	*3.4*	*-65.5*	*10.4*	*-40.7*	*7.7*	*179.5*	*3.6*	*-4.2*	*1.4*
n5-2	283	-179.9	4.8	-118.7	13.7	-50.6	13.0	179.6	4.7	-3.8	1.4
n5-2	*208*	*179.7*	*4.4*	*-116.8*	*15.6*	*-50.4*	*12.4*	*179.9*	*4.7*	*-3.9*	*1.4*
n5-3	219	-175.4	5.5	-67.1	14.9	-34.3	13.4	179.7	5.3	-2.9	1.5
n5-3	*153*	*-175.2*	*5.5*	*-66.3*	*13.8*	*-35.4*	*12.6*	*-179.4*	*4.8*	*-2.9*	*1.5*
n5-4	190	179.8	3.7	-63.0	6.0	-41.6	6.8	178.3	3.2	-1.3	0.7
n5-4	*138*	*-179.7*	*3.8*	*-63.1*	*6.0*	*-41.9*	*6.6*	*178.2*	*3.6*	*-1.3*	*0.8*
n5-5	57	-175.0	5.3	-126.4	14.9	-35.7	14.4	-177.6	5.6	-3.5	1.3
n5-5	*33*	*-173.0*	*4.5*	*-125.2*	*15.5*	*-33.2*	*14.6*	*-176.7*	*5.6*	*-3.6*	*1.3*
n5-6	52	177.5	3.9	64.2	12.1	40.6	11.4	179.5	3.8	-1.8	0.9
n5-6	*37*	*177.9*	*4.4*	*62.4*	*9.8*	*40.8*	*11.2*	*179.2*	*4.7*	*-1.5*	*0.7*
n5-7	43	-179.9	5.0	-70.5	14.4	-34.0	12.2	-179.3	5.0	-4.0	1.6
n5-7	*25*	*-179.6*	*8.0*	*-72.4*	*17.2*	*-33.5*	*11.5*	*178.3*	*6.8*	*-3.5*	*1.9*
n5-8	28	176.6	4.7	112.5	13.5	29.4	13.7	-179.3	4.0	-2.5	1.1
n5-8	*19*	*176.6*	*4.6*	*109.7*	*15.3*	*34.1*	*11.7*	*-179.6*	*3.5*	*-2.8*	*1.1*
n5-9	31	177.9	4.9	69.0	10.5	35.1	13.9	179.0	4.2	-3.3	1.4
n5-9	*27*	*177.5*	*4.7*	*71.4*	*12.2*	*30.9*	*19.0*	*179.7*	*4.6*	*-3.0*	*1.3*
n5-10	20	176.6	3.6	84.2	25.3	15.5	20.3	178.6	3.0	-2.4	1.3
n5-10	*14*	*176.1*	*3.6*	*87.1*	*22.7*	*13.0*	*17.9*	*178.3*	*4.1*	*-2.1*	*1.0*
n5-11	21	-176.9	5.3	-97.8	20.0	-36.6	12.1	-175.8	5.7	-3.0	1.3
n5-11	*11*	*-178.4*	*4.8*	*-91.0*	*16.8*	*-37.5*	*11.1*	*-175.3*	*5.4*	*-2.8*	*1.2*
n5-12	17	-170.0	5.7	-135.2	13.1	1.3	11.9	-178.9	3.5	-3.4	1.4
n5-12	*9*	*-173.0*	*3.3*	*-134.2*	*12.2*	*4.4*	*11.0*	*-179.1*	*2.7*	*-3.0*	*1.0*
n5-13	17	-178.6	3.5	56.8	7.6	36.5	7.0	177.3	4.1	-4.2	1.2
n5-13	*14*	*179.8*	*4.5*	*58.9*	*7.5*	*36.4*	*6.3*	*176.7*	*4.0*	*-4.5*	*1.1*
n5-14	9	-172.4	4.8	-141.9	6.8	10.5	4.7	179.6	3.3	-4.3	1.4
n5-14	*3*	*-174.6*	*0.4*	*-135.5*	*2.0*	*14.3*	*2.1*	*178.1*	*4.6*	*-5.0*	*1.4*

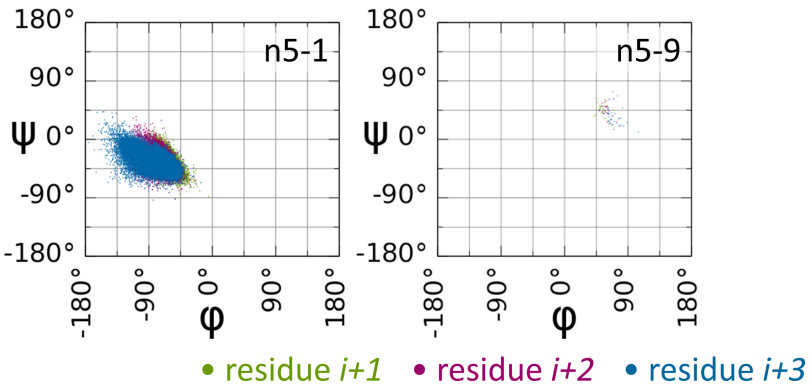


Figure 4.3.7. Ramachandran plots for the normal 5 turn types used for the α-helix assignment. Turn type n5-1 (left panel) is found as repetitive motif in right-handed α-helices, while successive turn structures of type n5-9 (right panel) are characteristic of left-handed α-helices.

Normal 6 Turns

The difference in the number of distinct turn types for normal 6 turns results from the differently sized datasets used for ESOM training. As observable in the corresponding U-Matrix (Figure 4.3.8), most of the new types are assigned to only a small number of turns in the training set (turn types n6-7 to n6-10, and n6-12 to n6-19). These were not classified before as their number was too small in the former ESOM-based classification. Their clustering became feasible due to the higher number of turns in the training set and a multiplier of 100 used for the ESOM training which enabled a clear differentiation between these types.

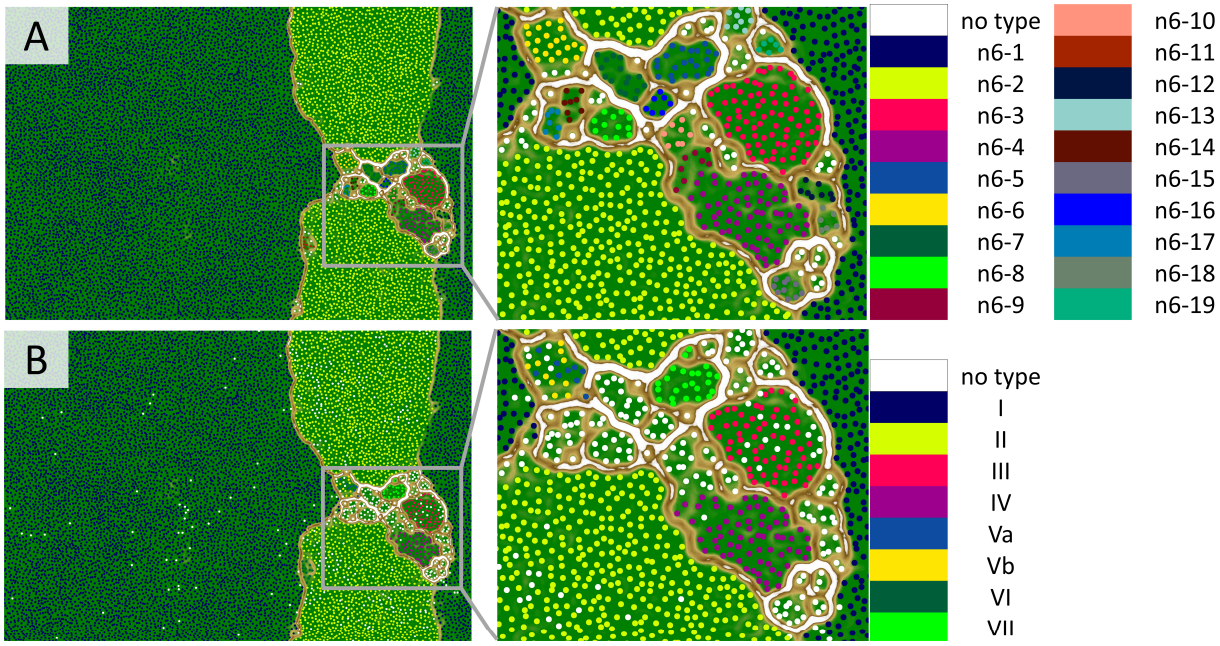


Figure 4.3.8. Normal 6 turn clustering. The U-Matrix for normal 6 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown.

Despite these differences, the Secbase-defined normal 6 turn types are closely related to the new highest occupied types: n6-1 and I, n6-2 and II, n6-3 and III, n6-4 and IV, n6-5 and VII, n6-6 and Va/Vb, and n6-11 and VI. However, unassigned turns for the former classification are detected in nearly all of the new clusters.

The types n6-1 and n6-2 (a type which is characterised by identical dihedral angles of the successive residues) differ mainly in the dihedral angles of residue *i+4*. Turn types n6-8 and n6-15 differ in the dihedral angle of residues *i+4* or *i+1*. All four turn types are characterised by a stable hydrogen bond between residue *i* and *i+5* (see Table 4.3.4). The turns of type n6-2 show dihedral angles preferentially found within π-helices (4.4_{16}-helices; φ = -57°, ψ = -70°). Their occurrence probably arises from one residue insertions during evolution which lead to distorted α-helical conformations.[97] These turns are often observed at varying positions of α-helices and are also referred to as α-bulges.[426] They can be used for the assignment of π-helical regions inside other helix types, but also of purely π-helical protein backbone segments. The left-handed counterpart of the n6-2 turn (φ = +57°, ψ = +70°) was not found in the dataset. Turns of type n6-1 (formerly normal 6 I) were previously used to assign the caps of π-helices at the C-terminus.[106]

Table 4.3.4. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of normal 6 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}		φ_{i+2}		ψ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
n6-1	8,430	179.1	3.5	-63.8	5.6	-41.7	7.7	180.0	3.2	-64.0	7.2	-31.1	9.8
n6-1	*5,941*	*179.1*	*3.4*	*-63.7*	*5.6*	*-41.7*	*7.8*	*179.9*	*3.1*	*-64.0*	*7.2*	*-31.0*	*9.6*
n6-2	2,157	-178.7	4.3	-66.5	9.2	-40.9	10.8	-176.8	4.4	-81.7	13.6	-35.7	17.6
n6-2	*1,563*	*-178.6*	*4.2*	*-66.6*	*9.4*	*-41.1*	*11.1*	*-176.9*	*4.3*	*-81.5*	*13.5*	*-36.5*	*17.5*
n6-3	80	178.9	4.8	54.8	6.9	49.7	10.8	179.1	3.3	59.7	7.8	36.9	12.4
n6-3	*51*	*179.1*	*4.5*	*55.8*	*5.8*	*50.0*	*10.4*	*179.7*	*3.3*	*59.6*	*7.4*	*38.3*	*12.5*
n6-4	67	-177.1	5.3	-65.6	8.4	126.7	6.9	-179.0	3.9	79.1	12.9	0.3	13.9
n6-4	*45*	*-177.4*	*5.7*	*-64.5*	*6.7*	*126.4*	*7.0*	*-178.5*	*3.7*	*80.7*	*10.6*	*-1.4*	*13.1*
n6-5	22	179.3	4.7	-97.1	14.1	-28.6	20.3	-179.7	3.3	-135.2	8.5	101.5	11.7
n6-5	*14*	*-179.6*	*3.4*	*-98.8*	*14.9*	*-31.0*	*15.5*	*179.2*	*5.0*	*-134.1*	*9.4*	*101.8*	*13.8*
n6-6	22	175.5	7.3	77.0	9.4	-178.2	11.5	-177.5	4.9	-63.2	7.3	-29.0	9.8
n6-6	*15*	*175.7*	*8.8*	*77.4*	*7.0*	*-179.5*	*9.9*	*-176.8*	*5.9*	*-62.2*	*6.9*	*-28.9*	*9.1*
n6-7	16	172.6	4.6	101.9	12.1	173.6	11.3	175.1	4.7	-130.2	6.0	107.8	7.2
n6-7	*16*	*173.6*	*4.6*	*98.3*	*10.7*	*176.8*	*10.7*	*177.7*	*3.6*	*-131.4*	*7.9*	*106.0*	*8.8*
n6-8	12	-178.6	4.2	-71.3	9.6	-28.1	12.0	-176.4	4.5	-79.2	11.1	-25.8	15.2
n6-8	*7*	*-177.2*	*5.5*	*-72.2*	*8.0*	*-27.9*	*11.4*	*-175.7*	*6.0*	*-79.6*	*7.8*	*-20.8*	*12.0*
n6-9	8	180.0	3.4	-74.2	13.2	123.1	21.7	-177.6	2.8	96.2	28.1	-57.9	10.2
n6-9	*7*	*-178.8*	*2.7*	*-74.8*	*13.8*	*117.0*	*19.7*	*-178.7*	*1.5*	*98.6*	*28.7*	*-55.0*	*6.4*
n6-10	8	179.1	7.6	-61.2	4.7	145.3	16.0	-178.3	5.9	96.1	17.3	-135.8	11.6
n6-10	*5*	*178.9*	*6.1*	*-58.1*	*4.0*	*142.4*	*6.9*	*179.4*	*2.6*	*97.6*	*10.1*	*-135.5*	*10.1*
n6-11	9	-177.8	4.2	-79.5	9.2	-10.5	10.5	-179.5	4.2	-107.7	13.4	42.3	10.9
n6-11	*7*	*-177.5*	*2.9*	*-79.3*	*13.3*	*-4.5*	*12.6*	*-179.7*	*4.2*	*-109.7*	*20.1*	*35.5*	*15.7*
n6-12	8	179.6	5.0	-60.7	4.5	129.4	2.7	-178.5	1.9	74.5	9.4	3.0	9.6
n6-12	*4*	*-179.5*	*4.3*	*-60.5*	*3.3*	*128.6*	*5.2*	*-178.1*	*3.3*	*68.2*	*13.5*	*10.7*	*8.5*
n6-13	9	-179.2	6.7	-92.2	8.5	-70.7	3.9	-176.5	4.2	-66.1	6.5	125.2	7.2
n6-13	*7*	*-176.6*	*5.9*	*-93.4*	*7.0*	*-69.2*	*5.0*	*-175.5*	*3.6*	*-64.6*	*7.3*	*121.7*	*7.1*
n6-14	8	-178.7	2.9	-61.8	4.9	-40.4	6.3	176.5	2.8	-89.3	8.4	-38.0	13.7
n6-14	*7*	*-178.2*	*4.5*	*-60.0*	*6.2*	*-42.7*	*4.3*	*177.2*	*3.3*	*-90.2*	*7.6*	*-37.4*	*14.5*
n6-15	10	-178.7	8.2	-118.3	13.5	128.5	8.3	3.7	4.8	-91.5	3.7	-4.9	11.8
n6-15	*8*	*177.6*	*5.9*	*-122.0*	*15.2*	*124.8*	*5.3*	*5.9*	*3.6*	*-92.6*	*2.5*	*-0.4*	*11.6*
n6-16	7	176.9	6.7	-122.4	10.5	163.9	6.3	176.6	3.4	52.1	5.1	63.6	7.5
n6-16	*2*	*168.9*	*4.3*	*-124.0*	*6.7*	*163.9*	*1.2*	*177.3*	*0.4*	*51.1*	*7.7*	*59.3*	*2.3*
n6-17	7	179.2	2.9	-68.3	6.9	-30.2	13.3	177.0	3.4	-84.3	14.4	-45.1	7.9
n6-17	*5*	*176.3*	*3.3*	*-63.5*	*7.4*	*-35.0*	*12.5*	*178.2*	*3.6*	*-78.8*	*11.0*	*-44.9*	*7.7*
n6-18	6	179.1	2.1	-75.0	6.7	158.2	6.6	-178.5	5.1	83.2	10.6	-55.0	14.8
n6-18	*3*	*-179.2*	*1.3*	*-71.0*	*2.8*	*156.4*	*5.3*	*-172.7*	*4.4*	*88.3*	*9.6*	*-54.4*	*14.6*
n6-19	6	176.4	5.2	-65.0	5.0	-24.9	12.1	177.7	3.4	-74.0	5.2	99.2	10.2
n6-19	*5*	*176.0*	*5.2*	*-64.9*	*8.2*	*-32.1*	*19.2*	*175.8*	*4.9*	*-70.9*	*8.0*	*107.4*	*17.1*

Table 4.3.4 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of normal 6 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+2}		φ_{i+3}		ψ_{i+3}		ω_{i+3}		φ_{i+4}		ψ_{i+4}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
n6-1	8,430	179.8	3.7	-91.3	12.5	1.5	11.9	178.8	4.6	74.7	19.2	26.9	16.8
n6-1	*5,941*	*179.7*	*3.7*	*-91.2*	*12.5*	*1.6*	*11.9*	*178.8*	*4.6*	*75.0*	*19.4*	*26.7*	*16.9*
n6-2	2,157	-174.9	5.1	-99.9	21.1	-44.8	23.7	-175.9	5.0	-91.7	25.5	-46.5	14.3
n6-2	*1,563*	*-174.9*	*5.0*	*-99.0*	*21.5*	*-44.7*	*23.3*	*-176.0*	*4.9*	*-91.4*	*25.6*	*-46.7*	*14.2*
n6-3	80	178.8	3.5	83.7	16.0	4.7	17.7	-179.4	5.2	-82.4	17.8	-21.6	13.9
n6-3	*51*	*178.6*	*3.4*	*82.3*	*16.4*	*5.7*	*18.1*	*-179.5*	*5.1*	*-84.1*	*19.2*	*-22.3*	*14.4*
n6-4	67	177.9	4.9	-119.0	11.9	-70.6	18.7	177.9	3.6	-91.5	20.0	-26.6	14.8
n6-4	*45*	*178.6*	*5.4*	*-118.3*	*14.2*	*-69.4*	*18.1*	*177.6*	*5.1*	*-92.1*	*20.1*	*-26.5*	*13.6*
n6-5	22	-4.2	5.2	-74.3	6.1	166.0	8.4	-178.6	4.0	-71.5	7.9	114.5	12.1
n6-5	*14*	*-2.8*	*4.7*	*-73.7*	*8.0*	*165.7*	*9.6*	*-177.7*	*3.2*	*-70.9*	*9.0*	*117.1*	*13.3*
n6-6	22	179.5	4.2	-91.4	12.9	0.9	13.0	178.8	4.7	53.7	9.8	56.3	7.9
n6-6	*15*	*179.0*	*3.5*	*-88.8*	*12.6*	*-3.8*	*13.5*	*179.6*	*4.2*	*56.4*	*11.2*	*56.5*	*9.0*
n6-7	16	-2.0	4.2	-72.3	4.9	161.4	5.9	-176.1	3.0	-65.0	6.7	134.1	5.7
n6-7	*16*	*0.1*	*4.2*	*-73.3*	*6.8*	*159.5*	*6.8*	*-175.6*	*3.2*	*-65.2*	*7.7*	*135.0*	*9.0*
n6-8	12	-177.0	5.0	-67.1	11.5	-54.3	2.2	-7.8	5.8	-91.2	4.4	156.2	8.9
n6-8	*7*	*-176.6*	*6.7*	*-71.1*	*8.5*	*-57.4*	*2.8*	*-4.6*	*3.0*	*-91.2*	*3.5*	*152.9*	*8.7*
n6-9	8	-177.8	4.0	-72.6	13.6	-37.1	11.7	-180.0	6.3	-108.8	21.5	-45.3	9.3
n6-9	*7*	*-176.7*	*4.9*	*-73.7*	*11.7*	*-39.6*	*14.5*	*179.7*	*5.1*	*-106.5*	*23.7*	*-44.9*	*9.2*
n6-10	8	-172.0	3.7	-59.7	5.5	-38.2	8.8	-178.9	4.8	-62.4	5.5	-51.2	7.6
n6-10	*5*	*-174.1*	*2.5*	*-56.1*	*6.5*	*-41.3*	*10.0*	*-178.8*	*1.9*	*-62.8*	*2.3*	*-49.7*	*2.8*
n6-11	9	173.8	7.8	125.2	21.8	-29.6	15.2	-178.1	4.6	-65.9	11.6	-39.4	12.7
n6-11	*7*	*175.8*	*5.3*	*127.5*	*24.0*	*-27.9*	*16.8*	*-179.4*	*3.7*	*-65.8*	*7.8*	*-41.9*	*9.6*
n6-12	8	-179.3	2.0	-131.0	11.2	29.2	16.5	178.5	4.3	115.7	17.3	-16.6	16.2
n6-12	*4*	*-180.0*	*2.1*	*-135.5*	*14.2*	*33.5*	*24.0*	*179.8*	*5.4*	*114.2*	*6.3*	*-12.9*	*13.4*
n6-13	9	-179.6	2.4	74.1	11.0	7.0	10.0	179.0	3.9	-127.5	10.7	-91.4	12.9
n6-13	*7*	*-177.9*	*3.6*	*77.0*	*7.5*	*4.8*	*6.8*	*-178.0*	*5.0*	*-128.8*	*13.1*	*-85.5*	*10.9*
n6-14	8	-173.7	3.8	-122.3	11.5	89.9	9.0	2.7	6.4	-133.9	13.8	28.9	14.2
n6-14	*7*	*-173.3*	*3.1*	*-120.7*	*10.8*	*89.4*	*10.1*	*3.1*	*4.4*	*-132.5*	*14.9*	*29.1*	*15.5*
n6-15	10	-171.7	6.2	-58.7	8.5	-46.7	7.0	-178.4	3.9	-94.8	16.3	-42.2	9.5
n6-15	*8*	*-171.4*	*5.2*	*-58.0*	*4.9*	*-44.6*	*5.6*	*-176.6*	*3.4*	*-101.0*	*15.6*	*-40.7*	*9.9*
n6-16	7	2.2	4.8	-90.0	3.8	164.6	5.2	178.3	3.5	-64.6	7.5	124.8	7.2
n6-16	*2*	*5.0*	*3.4*	*-85.9*	*3.8*	*156.5*	*11.9*	*175.0*	*0.2*	*-68.8*	*2.9*	*116.3*	*4.5*
n6-17	7	177.2	3.0	-145.1	10.1	122.0	8.3	7.9	5.4	-84.0	8.3	-5.7	10.1
n6-17	*5*	*175.3*	*4.6*	*-146.9*	*10.5*	*124.6*	*6.1*	*7.3*	*1.9*	*-88.1*	*8.5*	*-1.0*	*10.6*
n6-18	6	174.5	6.2	-78.8	13.8	-19.2	7.7	177.7	4.3	65.0	11.7	41.6	18.8
n6-18	*3*	*168.0*	*1.8*	*-74.9*	*14.6*	*-20.3*	*11.4*	*-179.9*	*4.0*	*63.8*	*13.8*	*35.3*	*14.9*
n6-19	6	-178.6	2.2	88.6	15.2	7.2	8.1	-176.1	6.3	43.7	5.9	58.4	5.1
n6-19	*5*	*-178.9*	*1.0*	*87.9*	*16.5*	*5.5*	*8.5*	*-176.9*	*7.9*	*44.9*	*5.4*	*59.3*	*5.0*

Table 4.3.4 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of normal 6 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+4}		E_{hb} [kcal/mol]	
		∅	±	∅	±
n6-1	8,430	179.3	4.2	-4.0	1.3
n6-1	*5,941*	*179.4*	*4.2*	*-4.0*	*1.3*
n6-2	2,157	-177.1	5.1	-4.1	1.6
n6-2	*1,563*	*-177.1*	*5.0*	*-4.1*	*1.6*
n6-3	80	-179.9	4.8	-3.2	1.1
n6-3	*51*	*179.6*	*5.2*	*-3.3*	*1.1*
n6-4	67	-180.0	6.0	-2.9	1.3
n6-4	*45*	*-180.0*	*5.1*	*-2.6*	*1.1*
n6-5	22	180.0	4.9	-3.2	1.3
n6-5	*14*	*179.2*	*4.5*	*-2.8*	*1.1*
n6-6	22	178.8	4.4	-2.9	1.2
n6-6	*15*	*-179.5*	*3.5*	*-2.9*	*1.3*
n6-7	16	175.3	5.4	-3.5	1.1
n6-7	*16*	*175.4*	*6.1*	*-3.4*	*0.9*
n6-8	12	177.5	6.4	-3.2	1.2
n6-8	*7*	*175.3*	*4.6*	*-2.9*	*0.8*
n6-9	8	178.0	6.2	-3.9	2.1
n6-9	*7*	*178.8*	*6.6*	*-4.5*	*1.8*
n6-10	8	179.7	3.2	-2.7	1.1
n6-10	*5*	*178.6*	*3.7*	*-2.7*	*1.2*
n6-11	9	-179.9	2.7	-2.8	1.1
n6-11	*7*	*178.6*	*3.7*	*-2.4*	*1.0*
n6-12	8	177.9	6.0	-3.1	1.0
n6-12	*4*	*178.8*	*5.1*	*-3.0*	*1.9*
n6-13	9	178.5	3.4	-2.7	1.7
n6-13	*7*	*179.9*	*3.9*	*-2.2*	*1.6*
n6-14	8	-175.2	7.9	-4.1	1.5
n6-14	*7*	*-174.5*	*7.2*	*-3.8*	*1.4*
n6-15	10	-180.0	5.6	-3.8	1.4
n6-15	*8*	*178.9*	*5.8*	*-4.2*	*1.5*
n6-16	7	-179.7	1.8	-2.6	0.7
n6-16	*2*	*-179.1*	*1.9*	*-2.9*	*0.3*
n6-17	7	-176.2	6.6	-3.8	1.6
n6-17	*5*	*-171.8*	*7.0*	*-4.5*	*1.5*
n6-18	6	-178.9	0.9	-3.0	1.0
n6-18	*3*	*-179.7*	*2.0*	*-1.8*	*0.4*
n6-19	6	-173.4	6.1	-4.0	1.3
n6-19	*5*	*-172.7*	*9.1*	*-4.7*	*0.7*

The turn types n6-8 and n6-15 are commonly found as bends in disordered structural regions, but also at the termini of α-helices. The dihedral angles (see Figure 4.3.9) of the C-terminal residues of types n6-1 and n6-8 lie outside the typical range (Ccap), while the N-terminal residues of turns of type n6-15 lie outside the dihedral angle range for π-helices (Ncap). These turn types might be useful for the definition of novel helix-capping motifs.

Over- and underrepresented residues could be reliably identified for turns of types n6-1 to n6-4 (see Table A3.4). Gly residues frequently occur in all types except for type n6-2. The over- and underrepresented residues of the latter turn type are in accordance with the residue preferences in π-helices which were assigned by the geometry-based method ASSP[96] (overrepresented: Leu, Phe, Tyr, underrepresented: Ala, Gly, Pro, Ser). Intriguingly, this type is one of those with the highest standard deviations of the dihedral angles. For turn type n6-1, the most distinct residue preferences were found. This turn type is preferentially found at the C-terminus of helices underlining its unique impact on the helix capping.

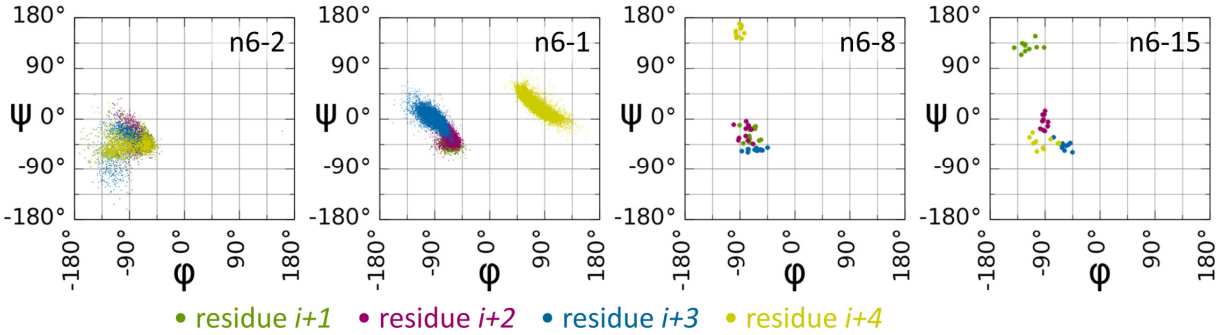


Figure 4.3.9. The dihedral angle ranges of selected normal 6 turn types. Ramachandran plots of the turn structures of normal 6 turn types that show dihedral angle ranges that are typical for π-helices (n6-2), helices at the C-terminus (n6-1 and n6-8), and helices at the N-terminus (n6-15).

Open Turns

The assigned open turns are not stabilised by hydrogen bonds between the terminal residues. Nonetheless, distinct types were previously classified for Cα-Cα distances between 4 and 10 Å.[278] Due to the high number of open turns that were detected, the maximum distance was decreased to 8 Å. The helix-capping procedure in SHAFT remained largely unaffected by this change (see Section 3.3.1). Turns corresponding to all previously defined turn types were identified. Due to the increased size of the maps applied for the new turn assignment procedure, the number of open turn types increased significantly (33 vs. 24 open 4 turn types, 56 vs. 22 open 5 turn types, and 81 vs. 28 open 6 turn types for the new and the Secbase-derived turn type assignment).

In SCOT, open turns are crucial for the assignment of appropriate helix termini as well as for the identification of PPII helices (open 4 turns) in proteins. This is further discussed in the following sections.

Open 4 Turns

The training set contains 99,883 open 4 turn structures (excluding turns that are completely spanned by normal and reverse turns) that were divided into 33 turn types. Again, we can identify general correspondences to the Secbase-assigned types as well as new clusters of yet unclassified types which result from the increased dataset size. A comparison of the U-Matrix without the best matches to that with the best matches coloured according to the original and new turn types (Figure 4.3.10) shows that the initial classification could be significantly improved. Differing assignments by the former classification in distant valleys of the U-Matrix point towards inadequacies of the initial turn clustering approach with Secbase.

Some Secbase-assigned type clusters are sparsely occupied since turns with Cα-Cα distances above 8 Å were excluded. The following correspondences to the former assignments were identified: o4-1 to VIIIa/VIIIb/XII4a/XII4b (the validity of their merging is verified by the small dihedral angle standard deviations of the new type), o4-2 to I/X, o4-3 to XI, o4-4 to VII2/VII3, o4-5 to IXX/XIII, o4-6 to IX, o4-7 to nt-5, o4-9 to nt-1/nt-3/nt-6, o4-10 to nt-2/nt4, o4-11 to VIb, o4-12/o4-16 to VII1a/VII1b, o4-13 to nt-7, o4-15 to I', and o4-17 to VIc. No further correspondences to other new types were found. The numerous missing assignments (no type) based on Secbase hint at the more comprehensive classification by the new ESOM.

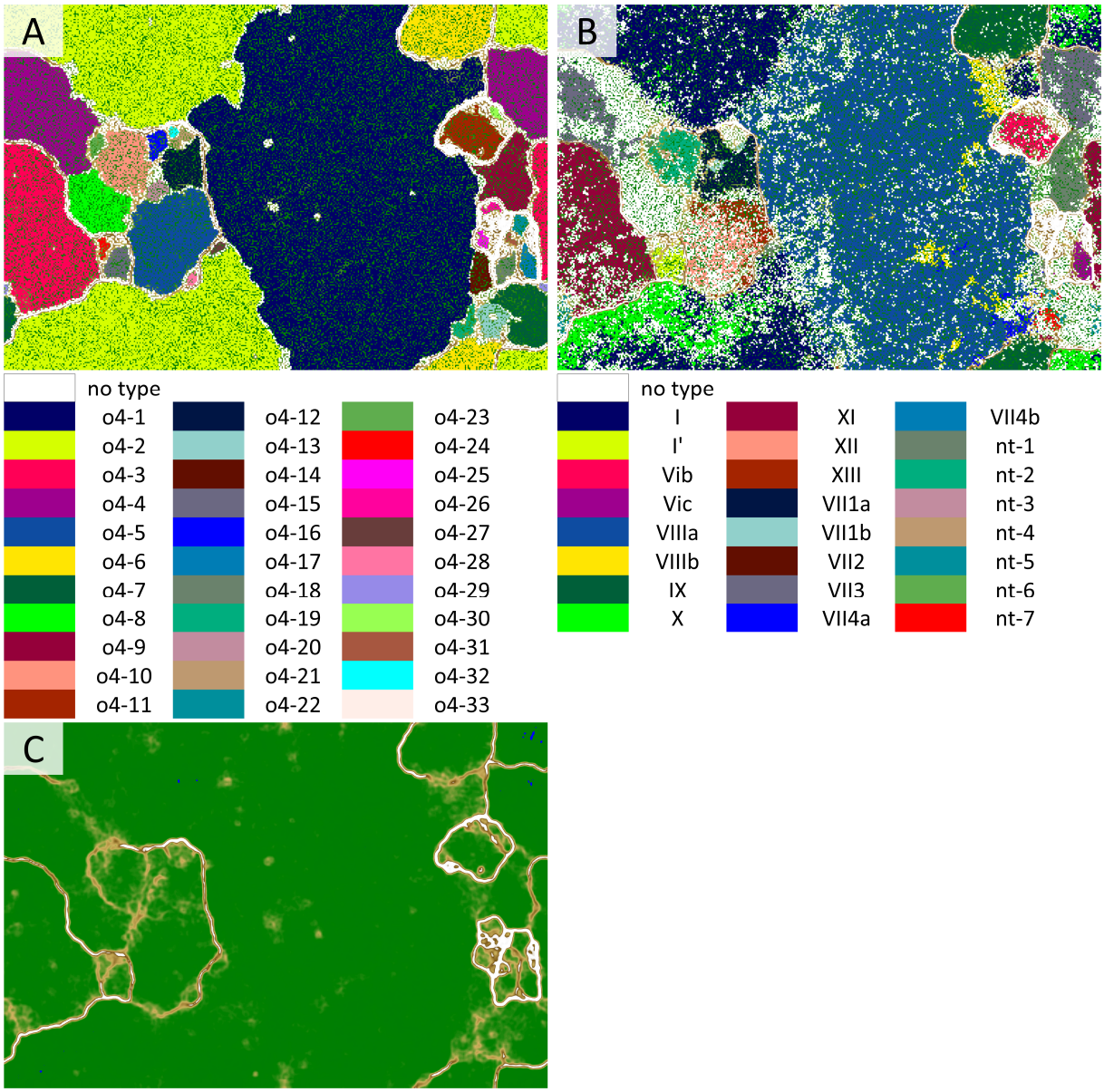


Figure 4.3.10. Open 4 turn clustering. The U-Matrix for open 4 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown. In (C), the U-Matrix without projected best matches is given.

The turns of type o4-1 are the most often occurring open 4 structures and are preferentially found at the N- and C-termini of helices and β-sheets. This is in line with the finding that Pro residues are preferentially found at positions *i* and *i+4* of this turn type. However, its meaning for the structure of proteins has not yet been explored.

As a general trend, the dihedral angles of the open turn types show higher standard deviations as compared to the hydrogen-bonded normal and reverse turns (Table 4.3.5, also Table A3.5). This is in line with the higher degree of freedom of the backbone structures of open turns which are not characterised by stabilising main-chain hydrogen bonds. Nonetheless, distinct types can be observed that mainly arise from stable side-chain contacts. The average Cα-Cα distances are similar for all turn types. For the turns of types o4-1 to o4-20, distinct residue preferences were found which are presented in Table A3.6 in Appendix 3.

The analysis of the Ramachandran plots for open 4 turn types (see Figure 4.3.11 and Table 4.3.5 for the corresponding dihedral angles) revealed some types with similar consecutive dihedral angles. They might become useful to assign SSEs. The turns of type o4-2 show the typical dihedral angle ranges of right-handed helices and are suitable for the extension of helices at the termini. They can be used for all helix types as their dihedral angle range covers the typical ranges of α-, 3_{10}-, and π-helices. Due to their repetitive ψ and φ angles, turns of type o4-15 are suitable for the extension of left-handed helices.

The PPII conformation of the main-chain (with average dihedral angles of φ = -75° and ψ = +145° as observed by Adzhubei and Sternberg[427]) is best represented by turns of type o4- A significant overrepresentation of Pro residues at all turn positions was observed (see Table A3.6). The fact that this backbone conformation is stabilised by interactions other than main-chain hydrogen bonds[428] underlines the necessity of using open turns for the assignment. The Cα-Cα distance for the classification of open turns (8 Å in these analyses) has to be increased to 10 Å to give consideration to the extended character of PPII helices.

The turn types o4-11 and o4-17 are also characterised by consecutive dihedral angles in the range of PPII helix angles (Table 4.3.5) and might be suitable for the identification of PPII helices. The overrepresented residue in these types is also Pro (and Tyr at position *i+1* in o4-11 turns). However, both types are characterised by a *cis* peptide bond which led to their exclusion for PPII assignment. As some studies suggest Lys as overrepresented residue in PPII conformations,[113] the dihedral angles of the turn types with a frequent occurrence of Lys residues were investigated. The only type that could be found was o4-1. However, due to the non-consecutive character of the underlying dihedral angles, this type should not be used for the assignment of PPII conformations in proteins.

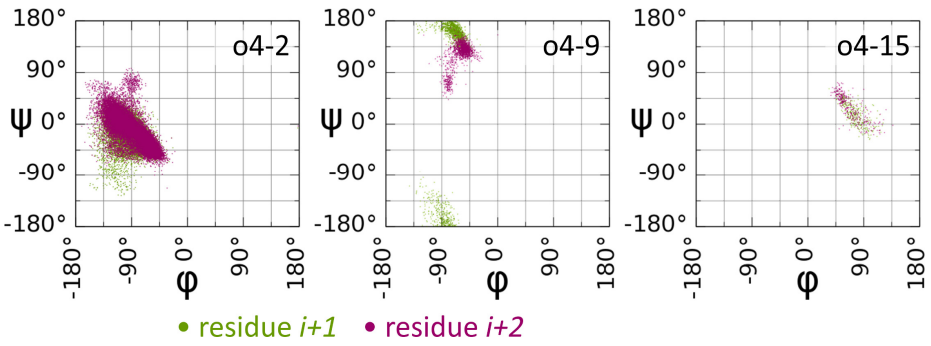


Figure 4.3.11. Ramachandran plots for open 4 turn types with an impact on helix assignment. Ramachandran plots of the open 4 turn types that show dihedral angle ranges which are typical for right-handed α-, 3_{10}-, and π-helices (o4-2), PPII helices (o4-9), and left-handed helices (o4-15).

Table 4.3.5. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 4 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}		φ_{i+2}		ψ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o4-1	39,622	180.0	5.2	-86.1	17.0	-21.0	18.3	180.0	5.2	-123.8	28.3	135.1	32.7
o4-1	*27,432*	*180.0*	*5.1*	*-86.1*	*17.1*	*-21.0*	*18.5*	*179.9*	*5.1*	*-123.8*	*28.4*	*135.1*	*32.8*
o4-2	24,781	-178.6	5.1	-87.2	19.5	-15.3	21.4	-176.4	5.2	-82.1	25.8	-17.8	27.6
o4-2	*17,524*	*-178.6*	*5.0*	*-87.0*	*19.5*	*-15.6*	*21.6*	*-176.5*	*5.1*	*-82.2*	*26.0*	*-17.7*	*27.7*
o4-3	6,989	179.0	5.6	-106.1	30.6	141.4	19.9	177.5	5.5	61.6	15.6	34.5	20.7
o4-3	*4,856*	*178.9*	*5.7*	*-106.2*	*30.7*	*141.2*	*20.1*	*177.8*	*5.5*	*61.3*	*15.8*	*34.8*	*20.5*
o4-4	6,971	175.5	6.2	-94.2	31.2	177.0	15.4	-179.2	4.9	-60.2	8.2	-30.0	12.8
o4-4	*4,841*	*175.6*	*6.3*	*-94.2*	*31.3*	*177.2*	*15.6*	*-179.2*	*4.9*	*-60.1*	*8.4*	*-30.1*	*12.8*
o4-5	3,398	179.2	4.6	80.2	15.5	5.9	18.7	-179.3	4.8	-93.5	24.7	-11.0	28.0
o4-5	*2,372*	*179.3*	*4.5*	*79.8*	*15.7*	*6.6*	*18.9*	*-179.3*	*4.9*	*-93.6*	*25.1*	*-11.0*	*28.2*
o4-6	3,014	-179.2	5.3	-97.5	19.9	2.2	16.2	179.7	5.0	78.2	18.4	14.7	21.1
o4-6	*2,105*	*-179.5*	*5.3*	*-97.1*	*20.2*	*2.0*	*16.1*	*179.7*	*4.9*	*77.7*	*18.0*	*14.9*	*21.0*
o4-7	1,442	176.2	5.8	-115.0	26.7	122.6	22.0	-177.4	6.0	62.7	12.1	-130.0	14.0
o4-7	*995*	*176.5*	*5.7*	*-114.8*	*27.2*	*123.4*	*22.9*	*-177.4*	*5.7*	*62.8*	*12.4*	*-130.1*	*14.2*
o4-8	1,333	179.1	6.0	84.7	18.6	-171.7	22.4	-175.9	5.2	-67.9	15.2	-26.6	15.3
o4-8	*947*	*179.0*	*6.0*	*84.2*	*18.7*	*-171.7*	*22.6*	*-176.0*	*5.2*	*-67.9*	*14.6*	*-26.7*	*14.8*
o4-9	1,429	173.4	6.4	-74.0	15.8	168.4	22.5	178.8	5.4	-58.8	12.3	123.2	20.5
o4-9	*968*	*173.5*	*6.6*	*-73.7*	*15.8*	*168.0*	*22.4*	*178.8*	*5.4*	*-58.8*	*12.0*	*122.5*	*20.9*
o4-10	1,146	-178.9	5.8	81.4	19.5	-153.7	25.2	-179.0	5.0	-78.3	17.9	121.5	24.6
o4-10	*849*	*-178.7*	*6.0*	*81.9*	*19.1*	*-153.5*	*26.6*	*-179.1*	*5.1*	*-77.7*	*18.0*	*121.4*	*24.3*
o4-11	1,051	178.4	6.0	-119.1	29.7	138.7	19.3	-1.7	5.1	-76.0	13.5	155.1	13.2
o4-11	*770*	*178.4*	*5.7*	*-118.4*	*29.4*	*138.4*	*19.6*	*-1.6*	*4.9*	*-77.1*	*14.8*	*155.4*	*13.7*
o4-12	742	-179.6	5.3	79.1	15.8	7.5	17.1	178.7	5.1	-150.3	14.8	158.0	21.7
o4-12	*510*	*-179.4*	*4.9*	*79.4*	*15.7*	*7.9*	*17.0*	*178.7*	*4.9*	*-149.9*	*15.0*	*158.5*	*22.2*
o4-13	337	177.5	5.3	-98.6	14.6	105.0	20.1	-177.2	4.7	91.6	16.5	148.5	34.5
o4-13	*247*	*177.9*	*5.1*	*-99.9*	*15.5*	*105.4*	*20.2*	*-176.9*	*5.0*	*92.3*	*15.9*	*146.5*	*35.1*
o4-14	339	-179.9	5.3	-87.3	8.7	49.9	13.8	179.2	5.7	-158.2	13.8	160.3	18.1
o4-14	*240*	*-179.4*	*4.9*	*-88.3*	*8.7*	*49.6*	*13.7*	*179.2*	*6.0*	*-157.3*	*13.0*	*160.9*	*18.6*
o4-15	284	178.0	8.0	76.2	16.1	17.7	19.1	176.4	5.4	68.4	19.5	30.2	22.3
o4-15	*192*	*178.3*	*4.8*	*75.2*	*17.1*	*18.9*	*20.5*	*176.9*	*5.4*	*68.2*	*19.5*	*32.0*	*22.7*
o4-16	206	177.1	6.0	78.5	9.0	-55.9	15.9	179.3	5.0	-146.5	14.6	152.5	16.2
o4-16	*138*	*177.5*	*6.2*	*79.5*	*10.7*	*-55.4*	*15.9*	*178.8*	*4.8*	*-146.4*	*14.5*	*151.2*	*17.7*
o4-17	208	-1.8	4.8	-67.7	6.4	155.0	13.4	-179.2	4.7	-69.8	12.1	133.1	23.2
o4-17	*147*	*-2.0*	*4.3*	*-67.9*	*10.0*	*155.5*	*12.6*	*-179.3*	*5.1*	*-68.7*	*12.3*	*133.0*	*24.8*
o4-18	189	5.6	5.8	-87.0	9.5	0.4	13.9	178.7	5.8	-90.9	30.0	147.9	15.7
o4-18	*130*	*5.6*	*5.5*	*-85.8*	*9.6*	*-0.4*	*14.5*	*179.6*	*5.9*	*-93.8*	*30.2*	*147.3*	*15.9*
o4-19	149	178.4	4.8	-92.7	15.8	4.1	15.0	-179.9	6.2	120.5	20.7	151.1	18.5
o4-19	*103*	*178.3*	*5.2*	*-92.0*	*15.9*	*4.1*	*15.7*	*-179.8*	*6.5*	*122.3*	*19.8*	*152.7*	*18.7*
o4-20	166	-179.3	4.6	88.7	10.4	-9.4	10.8	177.9	4.8	-101.2	13.7	91.4	18.6
o4-20	*124*	*-178.7*	*4.6*	*86.3*	*10.2*	*-8.0*	*9.9*	*177.0*	*6.4*	*-104.7*	*14.0*	*95.0*	*18.9*
o4-21	84	-178.9	5.1	72.4	16.6	23.2	21.2	-179.6	4.3	145.5	24.8	-170.6	15.3
o4-21	*59*	*-179.3*	*4.4*	*72.0*	*15.4*	*23.1*	*19.6*	*-179.1*	*4.3*	*146.9*	*21.8*	*-170.2*	*17.6*
o4-22	124	-2.1	5.5	-73.2	7.9	150.6	12.9	-178.0	4.9	-64.5	9.9	-32.2	9.4
o4-22	*89*	*-1.9*	*6.2*	*-74.1*	*7.8*	*149.9*	*13.5*	*-177.8*	*5.5*	*-63.8*	*8.8*	*-32.7*	*10.3*

Table 4.3.5 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 4 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+2}		Cα-Cα [Å]	
		∅	±	∅	±
o4-1	39,622	179.4	6.0	7.1	0.7
o4-1	*27,432*	*179.5*	*5.9*	*7.1*	*0.7*
o4-2	24,781	-179.0	5.0	6.4	0.9
o4-2	*17,524*	*-179.1*	*5.0*	*6.4*	*0.9*
o4-3	6,989	179.4	5.0	6.4	0.7
o4-3	*4,856*	*179.4*	*5.1*	*6.4*	*0.7*
o4-4	6,971	179.3	4.1	7.7	0.3
o4-4	*4,841*	*179.4*	*4.4*	*7.7*	*0.2*
o4-5	3,398	179.8	5.4	7.5	0.5
o4-5	*2,372*	*179.8*	*5.4*	*7.5*	*0.5*
o4-6	3,014	179.5	4.8	7.6	0.4
o4-6	*2,105*	*179.5*	*4.7*	*7.6*	*0.4*
o4-7	1,442	-179.7	4.2	7.2	0.5
o4-7	*995*	*-179.8*	*4.2*	*7.3*	*0.5*
o4-8	1,333	-179.8	4.6	7.1	0.8
o4-8	*947*	*-179.5*	*4.7*	*7.1*	*0.8*
o4-9	1,429	-178.1	5.1	7.8	0.2
o4-9	*968*	*-178.0*	*5.0*	*7.8*	*0.2*
o4-10	1,146	-179.0	5.7	7.3	0.6
o4-10	*849*	*-179.1*	*5.9*	*7.3*	*0.6*
o4-11	1,051	178.9	6.5	6.3	0.9
o4-11	*770*	*179.1*	*6.4*	*6.3*	*0.9*
o4-12	742	178.3	5.7	7.7	0.3
o4-12	*510*	*178.3*	*5.6*	*7.7*	*0.3*
o4-13	337	-178.6	5.6	7.3	0.6
o4-13	*247*	*-178.8*	*5.5*	*7.3*	*0.6*
o4-14	339	178.1	5.9	7.7	0.2
o4-14	*240*	*178.7*	*6.3*	*7.7*	*0.3*
o4-15	284	-179.5	5.4	6.5	0.8
o4-15	*192*	*-179.5*	*5.6*	*6.4*	*0.8*
o4-16	206	177.7	5.8	7.2	0.5
o4-16	*138*	*178.1*	*5.5*	*7.2*	*0.5*
o4-17	208	178.7	5.5	7.7	0.2
o4-17	*147*	*178.7*	*5.0*	*7.7*	*0.2*
o4-18	189	178.0	6.1	6.7	0.6
o4-18	*130*	*178.0*	*5.7*	*6.7*	*0.6*
o4-19	149	-180.0	5.1	7.7	0.3
o4-19	*103*	*180.0*	*6.1*	*7.7*	*0.3*
o4-20	166	-175.2	5.3	7.7	0.2
o4-20	*124*	*-176.0*	*5.2*	*7.7*	*0.3*
o4-21	84	-179.1	4.8	7.2	0.5
o4-21	*59*	*-179.3*	*5.1*	*7.2*	*0.5*
o4-22	124	-179.9	4.4	7.7	0.2
o4-22	*89*	*179.9*	*4.1*	*7.7*	*0.2*

Open 5 Turns

Compared to the 22 Secbase-assigned open 5 turn types, the new classification leads to a much higher number of turn types. Again, the standard deviations of the dihedral angles are higher than those for the hydrogen-bonded turns. Nevertheless, a clear differentiation between distinct types was possible (Figure 4.3.12). Although correspondences to formerly assigned open 5 turn types were found (o5-1 to II, o5-2 to Ia/Ib, o5-3/o5-5/o5-19 to IV/VI, o5-4 to IXa/IXb, o5-6 to XI, o5-7 to V/VII/VIII, o5-8 to XII/XIII, o5-9 to XVI, o5-11 to XIV, o5-15 to III, o5-17 to X, o5-20 to XVII, o5-22 to XVa/XVb, and o5-23 to XVIII), the classification by Secbase shows major limitations. Large clusters in the U-Matrix (e.g., o5-10 and o5-12) remained unclassified. In some cases, distinct Secbase-assigned types cluster in different valleys of the U-Matrix (e.g., types IV and VI). Only four turns of type kink-1 were found (now type o5-52). Although the new assignment captures a very high number of types, it is better suited to differentiate between turns with distinct dihedral angle ranges (Table 4.3.6).

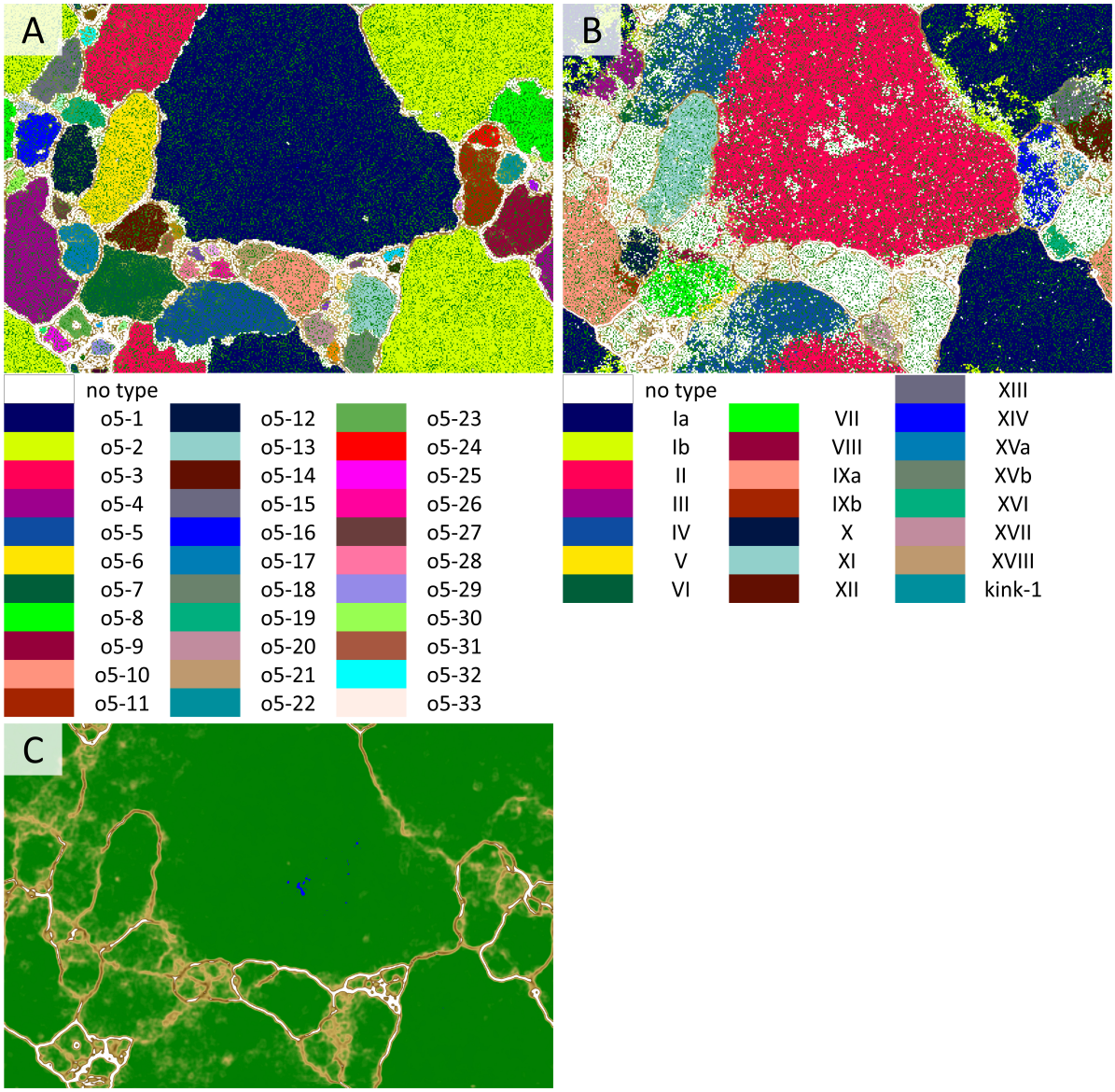


Figure 4.3.12. Open 5 turn clustering. The U-Matrix for open 5 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown. The legend in (A) only includes the 33 highest occupied turn types. In (C), the U-Matrix without the projected best matches is given.

Table 4.3.6. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 5 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}		φ_{i+2}		ψ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o5-1	41,710	-179.6	4.4	-65.7	13.3	-36.6	12.0	-179.1	4.0	-67.1	12.8	-27.8	13.4
o5-1	*29,289*	*-179.6*	*4.5*	*-65.8*	*13.5*	*-36.7*	*12.1*	*-179.2*	*4.0*	*-67.1*	*12.8*	*-27.8*	*13.4*
o5-2	28,500	177.7	5.8	-90.5	31.4	155.2	19.5	-179.1	5.0	-59.1	7.1	-33.3	11.3
o5-2	*19,671*	*177.8*	*5.7*	*-90.2*	*31.3*	*155.2*	*19.6*	*-179.1*	*4.9*	*-59.0*	*7.2*	*-33.4*	*11.3*
o5-3	5,945	-179.5	4.7	-67.5	11.9	-25.8	12.4	-177.8	4.7	-98.9	17.8	-6.8	17.5
o5-3	*4,122*	*-179.4*	*4.7*	*-67.7*	*11.8*	*-25.6*	*12.4*	*-177.9*	*4.8*	*-98.9*	*17.7*	*-6.4*	*17.7*
o5-4	4,660	178.5	5.7	-91.5	30.3	151.8	17.5	178.1	5.1	-56.4	7.4	134.8	9.0
o5-4	*3,245*	*178.6*	*5.6*	*-91.5*	*30.4*	*152.1*	*17.7*	*178.1*	*5.1*	*-56.5*	*7.6*	*135.0*	*9.1*
o5-5	3,296	178.8	4.6	-76.1	13.7	-29.7	14.6	-177.5	4.9	-91.4	18.7	-33.4	12.9
o5-5	*2,279*	*178.8*	*4.6*	*-75.8*	*13.6*	*-30.0*	*14.7*	*-177.4*	*4.9*	*-91.5*	*18.5*	*-33.1*	*13.0*
o5-6	3,659	-177.9	4.5	-73.3	16.5	-27.7	16.0	-178.4	5.0	-93.0	18.8	-7.4	16.2
o5-6	*2,622*	*-178.0*	*4.4*	*-72.9*	*16.3*	*-27.9*	*15.9*	*-178.5*	*5.0*	*-92.6*	*18.7*	*-7.5*	*16.1*
o5-7	2,631	179.1	5.4	-81.7	16.2	-16.2	23.4	177.9	5.4	-120.7	28.6	163.4	25.3
o5-7	*1,812*	*179.3*	*5.1*	*-81.9*	*16.3*	*-16.1*	*24.3*	*177.9*	*5.5*	*-121.3*	*28.5*	*163.7*	*24.6*
o5-8	2,027	178.3	6.0	77.3	21.9	-165.3	33.9	-176.8	4.8	-64.9	11.6	-26.2	16.2
o5-8	*1,463*	*178.4*	*5.9*	*76.4*	*21.8*	*-164.6*	*34.3*	*-176.8*	*4.7*	*-65.5*	*12.4*	*-26.0*	*16.3*
o5-9	2,226	178.0	5.2	-123.2	30.4	127.2	16.8	180.0	5.2	53.9	7.8	37.8	11.8
o5-9	*1,570*	*178.1*	*5.3*	*-122.9*	*30.1*	*127.3*	*17.8*	*-180.0*	*4.8*	*53.8*	*7.9*	*38.0*	*11.7*
o5-10	2,054	-179.8	4.7	52.9	6.2	39.7	9.0	177.6	3.4	78.5	10.9	4.2	13.3
o5-10	*1,453*	*-179.6*	*4.7*	*52.9*	*6.8*	*39.8*	*9.1*	*177.6*	*3.5*	*78.4*	*11.1*	*4.2*	*13.8*
o5-11	1,422	177.3	6.3	-119.4	24.4	123.9	26.0	-178.5	6.8	61.4	12.2	-128.2	13.6
o5-11	*985*	*177.4*	*6.2*	*-119.3*	*24.0*	*124.3*	*26.6*	*-178.5*	*6.6*	*60.9*	*11.9*	*-127.5*	*13.5*
o5-12	1,283	-179.1	4.6	-66.4	11.9	-29.7	12.5	-178.7	4.9	-93.9	15.4	-5.8	18.1
o5-12	*863*	*-179.1*	*4.8*	*-67.8*	*12.6*	*-29.0*	*12.7*	*-178.8*	*5.1*	*-93.4*	*15.8*	*-6.2*	*17.5*
o5-13	1,218	179.8	4.8	-61.3	15.8	129.9	10.2	179.0	3.8	76.1	13.4	7.1	17.6
o5-13	*842*	*179.7*	*4.8*	*-60.7*	*15.2*	*130.0*	*9.7*	*179.0*	*3.6*	*76.4*	*13.5*	*6.3*	*18.6*
o5-14	1,075	179.3	5.4	-88.1	18.8	-11.9	18.9	178.2	6.0	-105.3	28.3	173.8	19.7
o5-14	*718*	*179.5*	*5.5*	*-88.1*	*19.1*	*-11.3*	*19.0*	*178.3*	*6.0*	*-104.2*	*27.8*	*173.6*	*20.2*
o5-15	960	177.7	5.9	-94.4	31.9	163.5	19.6	179.1	4.9	-63.6	9.5	-33.2	13.7
o5-15	*665*	*177.9*	*5.9*	*-94.7*	*31.2*	*164.1*	*20.1*	*179.4*	*5.0*	*-63.4*	*9.9*	*-33.1*	*13.7*
o5-16	1,058	-176.0	5.8	60.4	9.1	-123.4	11.1	-179.8	3.5	-93.3	12.6	5.4	14.7
o5-16	*750*	*-176.2*	*5.6*	*60.3*	*9.0*	*-123.0*	*11.0*	*180.0*	*3.8*	*-93.6*	*13.2*	*4.6*	*15.3*
o5-17	851	-178.8	5.4	-88.8	18.7	-24.0	19.6	-178.5	4.9	-84.3	32.7	132.5	17.4
o5-17	*604*	*-178.7*	*5.5*	*-88.0*	*18.6*	*-23.7*	*19.9*	*-178.7*	*5.0*	*-85.4*	*33.0*	*132.6*	*18.1*
o5-18	688	-179.8	4.9	-56.6	7.6	130.3	8.0	179.8	3.5	82.5	13.3	-3.6	14.2
o5-18	*492*	*-179.8*	*5.6*	*-56.3*	*7.9*	*130.2*	*8.7*	*179.9*	*3.4*	*82.3*	*13.2*	*-3.2*	*14.3*
o5-19	453	179.4	5.2	-81.5	19.2	-30.8	19.6	-177.6	5.2	-87.5	15.6	-33.4	13.2
o5-19	*317*	*179.7*	*4.7*	*-82.8*	*18.7*	*-30.5*	*20.0*	*-178.2*	*4.8*	*-87.2*	*16.1*	*-33.3*	*13.7*
o5-20	414	-179.8	4.3	-91.4	17.1	-0.3	14.1	178.9	4.7	87.5	13.4	7.1	13.4
o5-20	*294*	*-179.9*	*4.1*	*-92.2*	*18.3*	*0.4*	*15.2*	*178.8*	*4.9*	*86.3*	*13.9*	*8.3*	*14.0*
o5-21	381	177.8	5.0	57.1	8.1	32.7	11.3	178.3	4.0	76.6	14.2	10.4	13.4
o5-21	*262*	*178.2*	*4.8*	*56.3*	*7.8*	*32.9*	*11.9*	*178.6*	*3.9*	*74.8*	*13.9*	*11.4*	*14.1*
o5-22	319	177.9	5.4	-96.5	27.4	144.7	19.2	179.2	5.1	82.5	17.1	-156.8	16.1
o5-22	*235*	*177.8*	*5.8*	*-97.9*	*26.5*	*145.6*	*19.1*	*179.2*	*4.9*	*82.4*	*16.4*	*-157.4*	*16.4*

Table 4.3.6 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 5 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+2}		φ_{i+3}		ψ_{i+3}		ω_{i+3}		Cα-Cα [Å]	
		∅	±	∅	±	∅	∅	∅	±	∅	±
o5-1	41,710	-179.3	4.3	-80.5	19.2	-14.8	19.3	-179.1	5.1	6.9	0.7
o5-1	*29,289*	*-179.4*	*4.3*	*-80.5*	*19.3*	*-14.9*	*19.5*	*-179.2*	*5.0*	*6.9*	*0.7*
o5-2	28,500	-180.0	3.6	-72.7	15.9	-23.7	20.3	179.3	4.2	6.6	0.7
o5-2	*19,671*	*-180.0*	*3.6*	*-72.5*	*15.9*	*-23.9*	*20.1*	*179.3*	*4.1*	*6.6*	*0.7*
o5-3	5,945	-178.6	5.3	-128.9	22.4	132.2	38.1	179.2	6.8	6.9	0.9
o5-3	*4,122*	*-178.6*	*5.2*	*-129.3*	*22.4*	*132.6*	*38.8*	*179.3*	*6.7*	*6.9*	*0.9*
o5-4	4,660	179.0	3.8	81.1	14.4	2.0	18.3	179.3	4.9	6.7	0.7
o5-4	*3,245*	*179.0*	*3.8*	*80.7*	*14.8*	*2.4*	*18.6*	*179.2*	*5.0*	*6.7*	*0.7*
o5-5	3,296	-177.7	5.3	-126.3	19.3	96.7	33.5	-177.9	5.9	5.8	1.1
o5-5	*2,279*	*-177.8*	*5.0*	*-127.0*	*18.9*	*95.4*	*33.6*	*-177.8*	*5.8*	*5.7*	*1.1*
o5-6	3,659	179.5	5.4	71.4	15.6	20.9	21.1	179.5	4.8	6.9	0.8
o5-6	*2,622*	*179.6*	*5.2*	*71.1*	*15.6*	*21.0*	*21.2*	*179.6*	*4.6*	*6.9*	*0.8*
o5-7	2,631	178.8	5.5	-64.6	13.6	132.4	20.5	-179.9	5.5	7.4	0.5
o5-7	*1,812*	*178.7*	*5.7*	*-64.7*	*13.3*	*132.2*	*20.9*	*-179.8*	*5.5*	*7.4*	*0.5*
o5-8	2,027	180.0	4.4	-71.9	16.1	-23.9	20.7	179.2	4.6	5.7	0.8
o5-8	*1,463*	*-179.8*	*4.4*	*-71.7*	*15.9*	*-24.5*	*20.1*	*179.3*	*4.5*	*5.7*	*0.8*
o5-9	2,226	177.9	3.8	74.9	12.1	8.2	15.7	179.9	4.4	5.5	0.5
o5-9	*1,570*	*177.9*	*3.8*	*75.0*	*12.2*	*8.3*	*15.8*	*179.9*	*4.5*	*5.5*	*0.6*
o5-10	2,054	179.7	4.1	-110.4	18.1	139.1	18.2	177.9	5.3	6.7	0.5
o5-10	*1,453*	*179.6*	*4.2*	*-110.1*	*17.9*	*139.2*	*18.2*	*178.0*	*5.2*	*6.7*	*0.5*
o5-11	1,422	-179.5	4.3	-87.5	17.8	-0.2	19.0	-179.3	5.2	5.8	0.8
o5-11	*985*	*-179.4*	*4.0*	*-86.6*	*17.8*	*-1.9*	*19.2*	*-179.3*	*6.3*	*5.7*	*0.8*
o5-12	1,283	-179.5	5.1	106.2	27.7	179.5	23.2	-178.2	4.9	6.7	0.8
o5-12	*863*	*-179.4*	*5.4*	*106.3*	*27.8*	*-180.0*	*23.6*	*-178.1*	*5.1*	*6.8*	*0.8*
o5-13	1,218	179.9	4.9	-125.2	20.7	153.1	26.2	177.9	6.3	7.3	0.6
o5-13	*842*	*179.9*	*4.7*	*-124.7*	*21.4*	*153.7*	*27.0*	*177.9*	*6.1*	*7.3*	*0.6*
o5-14	1,075	-179.0	5.2	-66.3	13.6	-24.1	19.1	179.8	5.2	7.5	0.5
o5-14	*718*	*-178.9*	*5.2*	*-66.2*	*13.5*	*-24.6*	*19.1*	*179.9*	*5.2*	*7.5*	*0.5*
o5-15	960	176.7	5.6	-116.0	19.6	107.2	27.8	-177.5	6.0	7.2	0.7
o5-15	*665*	*176.7*	*5.4*	*-115.3*	*20.4*	*108.0*	*28.2*	*-177.4*	*5.9*	*7.3*	*0.7*
o5-16	1,058	-178.3	5.1	-102.9	19.7	138.9	22.0	176.8	5.6	6.5	0.6
o5-16	*750*	*-178.4*	*5.3*	*-101.1*	*19.7*	*137.9*	*21.9*	*177.0*	*5.7*	*6.5*	*0.6*
o5-17	851	179.0	4.3	68.9	15.4	19.0	20.9	-179.5	5.0	6.9	0.9
o5-17	*604*	*179.2*	*4.4*	*69.1*	*15.4*	*19.4*	*21.4*	*-179.5*	*5.2*	*6.9*	*0.9*
o5-18	688	-178.4	5.5	-103.5	24.7	-1.9	29.4	-178.7	5.1	7.0	0.8
o5-18	*492*	*-178.2*	*5.5*	*-103.4*	*25.5*	*-2.4*	*31.0*	*-178.3*	*4.8*	*7.0*	*0.8*
o5-19	453	-178.3	5.2	-157.9	12.2	161.9	14.1	177.1	6.0	6.1	0.9
o5-19	*317*	*-178.8*	*5.1*	*-157.7*	*11.6*	*161.1*	*14.3*	*176.8*	*6.1*	*6.1*	*0.9*
o5-20	414	-179.9	4.9	-100.7	23.3	1.6	31.7	-179.9	5.3	7.6	0.5
o5-20	*294*	*-180.0*	*4.3*	*-101.3*	*24.2*	*0.3*	*31.6*	*-179.9*	*4.8*	*7.6*	*0.5*
o5-21	381	-177.8	4.9	-77.4	18.6	-22.8	20.2	179.8	5.2	6.6	0.7
o5-21	*262*	*-177.7*	*4.8*	*-76.9*	*18.4*	*-23.1*	*19.2*	*180.0*	*4.8*	*6.6*	*0.7*
o5-22	319	-178.0	5.2	-82.1	16.7	122.3	24.7	-179.3	6.1	7.1	0.7
o5-22	*235*	*-178.4*	*5.1*	*-81.9*	*16.9*	*119.7*	*24.0*	*-179.4*	*5.9*	*7.0*	*0.7*

Table 4.3.6 provides an overview of the most prominent open 5 turn types identified in this study. The Cα-Cα distance range is broader than that of the open 4 turn types. This finding and the larger number of distinct turn types can be attributed to the non-restrictiveness in the conformation of turns with a length of five residues.

An analysis of the Ramachandran plots for the turn types (see also Tables A3.7 and A3.8) shows that types o5-1 and o5-46 are of special interest for the extension of right-handed and left-handed helices, respectively. The turn type o5-23 is characterised by a consecutive pattern of dihedral angles similar to those of the PPII conformation. This type is indicative of the PPII helix and Pro residues are significantly overrepresented in these turns (see Table A3.9 in Appendix 3). The same holds true for the turn types o5-25 and o5-29, but their *cis* peptide bonds render them inapplicable for the assignment of PPII helices. All these types are situated in similar regions of the ESOM.

The turn type o5-1 (type II according to Secbase) was previously used to assign N-terminal helix caps. However, the regularity of the consecutive angles highlights its usability for the assignment of N- and C-terminal caps. According to the dihedral angles, the turn types o5-2 and o5-8 might also be used for the definition of Ncaps. The types o5-3, o5-5, o5-6, and o5-19 are applicable for the assignment of C-terminal helix caps. Most of them were formerly assigned to type IV which is used for the helix capping with SHAFT, but also to type VI. These suggestions for capping turns in proteins can be verified by analysing the overlaps between these turn types and regular SSEs in the training set (Table 4.3.7).

The high population of the o5-4 cluster is not explicable by the co-occurrence with helices in proteins. However, the dihedral angle ranges of the residues at positions *i+1* and *i+2* point towards their extended conformation. Pro and Gly residues are significantly overrepresented in turns of this type. Analysing the relationship between this turn type and SCOT-assigned SSEs, it is found that the turn type predominantly occurs in disordered regions outside the regular SSEs of the protein structures under investigation (Table 4.3.7).

Table 4.3.7. The relationship between the most frequent open 5 turns and helices. The fraction of turns of each type which is found completely inside helices (in) or completely outside helices (out) is presented. The fraction of turns overlapping by one, two, three, or four residues (O1, O2, O3, and O4) with the N- or C-terminus of helices shows typical relationships between the turn types and helices. A colour gradient from green (highest fraction) to yellow to red (lowest fraction) was applied.

Turn Type	in	out	Helix N-Terminus				Helix C-Terminus				Relationship to Helices
			O1	*O2*	*O3*	*O4*	*O1*	*O2*	*O3*	*O4*	
o5-1	0.74	0.02	0.02	0.01	0.01	0.13	0.02	0.01	0.01	0.10	Inside helices
o5-2	0.00	0.20	0.02	0.00	0.61	0.00	0.05	0.01	0.00	0.00	Ncap
o5-3	0.00	0.27	0.09	0.00	0.00	0.00	0.03	0.01	0.43	0.05	Ccap
o5-4	0.00	0.53	0.08	0.00	0.00	0.00	0.05	0.06	0.00	0.00	None
o5-5	0.00	0.23	0.17	0.00	0.00	0.00	0.01	0.02	0.38	0.26	Ccap
o5-6	0.00	0.24	0.03	0.00	0.00	0.00	0.01	0.01	0.68	0.00	Ccap
o5-8	0.00	0.22	0.05	0.04	0.54	0.04	0.21	0.00	0.00	0.00	Ncap
o5-19	0.00	0.41	0.06	0.00	0.00	0.00	0.02	0.04	0.28	0.00	Ccap

Open 6 Turns

The most significant improvement was observed for the classification of open 6 turns. While nearly half of the open 6 turn structures in the training set remained unclassified with the turn classification by Secbase, the new U-Matrix obtained from a larger dataset enables the classification of approximately 88% of all 83,711 turn structures. The manually created class mask assigns 81 clusters. Intriguingly, some highly occupied turn types were not assigned before (e.g., o6-3, o6-4, o6-5). Nevertheless, a comparison to the former assignment (Figure 4.3.13) enables the identification of some type correspondences: o6-1 to VII, o6-2 to III/XIV, o6-6 to I, o6-7 to II, o6-8 to VIIIa/VIIIb, o6-9 to XI, o6-12 to hook-1/hook-4, o6-14 to XIIIa/XIIIb, o6-15 to IXa/IXb, o6-18 to XVI/XVII, o6-19 to XV, o6-20/o6-67 to VI, o6-22 to XII, o6-23 to IV/V, o6-27 to Xa/Xb, o6-31 to XVIII, o6-33/o6-35 to hook-3, o6-38 to hook-6, o6-53/o6-62 to hook-5, and o6-57 to hook-2.

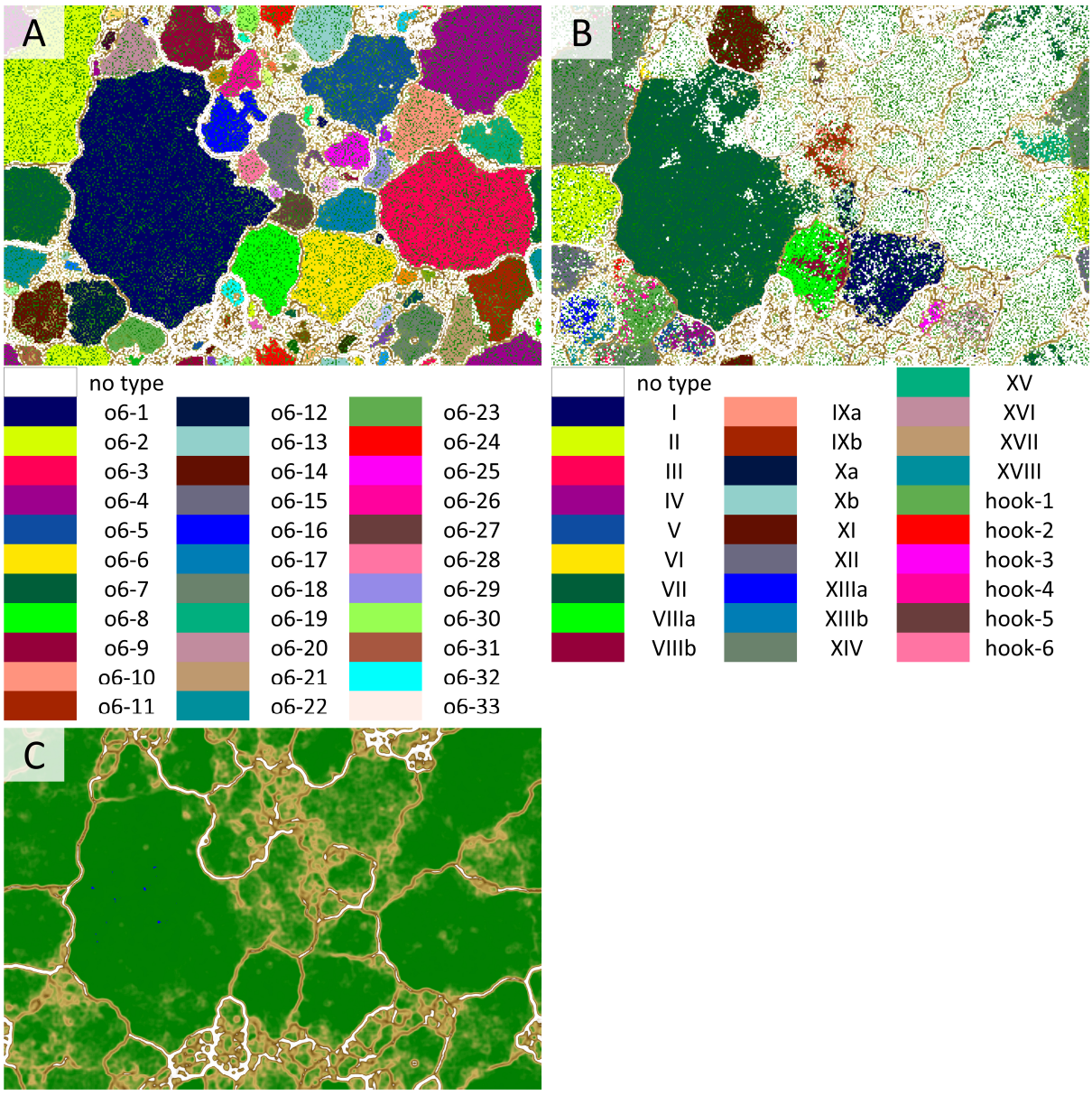


Figure 4.3.13. Open 6 turn clustering. The U-Matrix for open 6 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown. The legend in (A) only includes the 33 highest occupied turn types. In (C), the U-Matrix without the projected best matches is shown.

The turn types that correspond to more than one of the former types (o6-2, o6-12, o6-18, and o6-23) do not suffer from too high standard deviations and their dihedral angles cluster in distinct parts of the Ramachandran plot. The dihedral angle means and standard deviations of the different turn types (Table 4.3.8 and Tables A3.10-12) for the training and test set clearly show that the classification is robust. However, the standard deviations are higher as compared to the turns which are characterised by hydrogen bonds.

Of the highest occupied turn types, only type o6-4 is appropriate to extend helices at the more irregular termini which are not stabilised by main-chain hydrogen bonds. The corresponding φ and ψ angles are in a similar range and conform well with the general helix geometry. Based on the analysis of the Ramachandran plots, the turn types o6-1 and o6-20 can be used as Ccap motifs and are neighbouring types in the corresponding U-Matrix. Except for the residue at position *i+1*, the residues' dihedral angles are in the range of these of right-handed helices. The types o6-6, o6-10, and o6-17 are suitable for the assignment of helix Ncaps. Due to the similarities in the dihedral angles of the residues at the turn positions *i+2*, *i+3*, and *i+4*, they cluster in similar regions of the U-Matrix. However, the differences in the φ and ψ angles of the residue at the position *i+5* underpin the validity of their assignment to separate types. However, none of these turn types was used for the helix capping with SHAFT.

As a normal 6 turn type with mean dihedral angles corresponding to those of left-handed π-helices could not be detected within the clustered turn types, the Ramachandran plots of the open 6 turns were analysed to search for an open 6 turn type which might be indicative of this SSE. However, none of the types showed a consecutive pattern of ψ and φ angles that are characteristic of left-handed helices. As Kumar and Bansal proposed the occurrence of a left-handed π-helix in the crystal structure of split-Soret cytochrome C (PDB-ID 1h21, chain A),[122] the dihedral angles in this proposed helix from residues 77 to 81 were calculated. The φ angles range between 49° and 83.5°, while the ψ angles range between 20° and 91.2°. Besides the sparseness of its occurrence, it is challenging to assign this helix type due to these highly irregular angular characteristics. This might also explain the lack of normal and open turn types underlying left-handed π-helices in our dataset.

For the highest occupied turn types o6-2, o6-3, and o6-5, no repetitive sequence of dihedral angles is observed. They often occur between helices and strands. However, we find some distinct trends concerning their relationship to helices. Approximately 65% of the turns of type o6-2 overlap with the N-terminus of helices by three residues. Similarly, 67% and 43% of all o6-3 and o6-5 turns overlap by three residues with the C-terminus of helices. This hints at the potential involvement of these turn types in the stabilisation of the helix termini by distinct side-chain interactions. For the o6-3 turn type, 31% of the structures overlap with the N-terminus of strands by one residue. These relationships clearly indicate a potential impact of distinct open 6 turn types on the arrangement of SSEs in proteins. The three turn types are mainly occupied by charged residues. The turn types o6-2 and o6-5 additionally show a preference for Pro residues at distinct residue positions (see Table A3.13 for the conformational parameter of all residues in the highest occupied open 6 turn types).

Table 4.3.8. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}		φ_{i+2}		ψ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o6-1	19,044	177.4	5.8	-86.1	27.7	147.2	23.4	-178.8	5.1	-58.5	7.2	-37.5	9.4
o6-1	*13,293*	*177.5*	*5.6*	*-86.2*	*27.9*	*147.4*	*23.6*	*-178.8*	*5.1*	*-58.5*	*7.3*	*-37.6*	*9.6*
o6-2	7,143	-179.4	5.5	-90.6	19.1	-12.1	19.6	179.9	5.4	-101.3	33.4	143.7	24.2
o6-2	*4,924*	*-179.2*	*5.5*	*-90.6*	*19.2*	*-12.1*	*19.7*	*179.9*	*5.2*	*-101.3*	*33.6*	*143.7*	*24.2*
o6-3	7,070	-179.8	3.6	-63.6	7.5	-27.0	10.6	178.9	3.8	-90.1	12.2	2.3	11.5
o6-3	*4,936*	*-179.9*	*3.6*	*-63.5*	*7.3*	*-27.2*	*10.5*	*178.9*	*3.6*	*-90.0*	*12.2*	*2.3*	*11.6*
o6-4	6,539	179.9	4.2	-66.4	11.9	-40.6	12.1	-179.0	4.2	-73.0	17.9	-42.8	14.0
o6-4	*4,698*	*179.9*	*4.2*	*-66.5*	*11.9*	*-40.6*	*12.1*	*-179.1*	*4.2*	*-73.1*	*18.0*	*-43.1*	*14.2*
o6-5	2,742	179.4	5.0	-73.1	15.1	-26.2	14.9	-177.5	5.0	-98.5	18.3	-18.4	21.3
o6-5	*1,903*	*179.5*	*5.0*	*-73.4*	*15.2*	*-26.6*	*15.4*	*-177.6*	*5.0*	*-98.5*	*18.5*	*-18.1*	*21.1*
o6-6	2,988	-179.4	4.5	-67.2	10.5	-38.9	11.1	-178.1	3.9	-76.7	15.0	-33.3	13.7
o6-6	*2,160*	*-179.4*	*4.4*	*-67.2*	*10.6*	*-38.9*	*11.2*	*-178.2*	*3.9*	*-76.2*	*14.9*	*-33.2*	*14.0*
o6-7	2,291	-178.6	4.0	-64.4	9.7	-30.8	10.5	-179.8	3.9	-88.4	12.8	-1.8	11.8
o6-7	*1,632*	*-178.7*	*3.9*	*-64.3*	*10.1*	*-31.2*	*10.8*	*-179.8*	*3.9*	*-88.4*	*13.1*	*-1.7*	*12.0*
o6-8	2,136	177.6	5.7	-86.6	17.3	172.1	11.6	-178.2	4.3	-62.3	7.6	-23.2	11.0
o6-8	*1,446*	*177.7*	*5.6*	*-86.6*	*17.3*	*172.0*	*11.7*	*-178.3*	*4.2*	*-62.1*	*7.8*	*-23.3*	*11.1*
o6-9	1,877	177.6	5.2	-126.1	29.0	126.3	14.5	-179.7	4.5	54.2	9.4	38.2	12.1
o6-9	*1,306*	*177.7*	*5.2*	*-125.9*	*28.9*	*126.5*	*15.6*	*-179.6*	*4.7*	*54.0*	*9.6*	*38.4*	*12.2*
o6-10	1,566	-179.2	4.1	-64.2	7.5	-30.1	10.7	178.2	3.9	-72.8	11.1	-34.3	11.9
o6-10	*1,111*	*-179.3*	*4.0*	*-64.5*	*7.4*	*-30.1*	*10.4*	*178.2*	*4.0*	*-72.9*	*11.1*	*-34.3*	*12.0*
o6-11	1,438	178.5	4.5	-67.6	9.7	-32.9	10.3	-178.6	4.1	-90.7	16.6	-25.4	17.5
o6-11	*1,045*	*178.6*	*4.6*	*-67.9*	*9.8*	*-32.4*	*10.8*	*-178.6*	*4.0*	*-91.4*	*17.1*	*-25.4*	*18.0*
o6-12	1,114	-179.6	5.9	-93.3	23.1	91.4	29.6	179.2	5.9	-140.4	27.0	165.9	20.1
o6-12	*777*	*-179.5*	*5.4*	*-92.9*	*22.2*	*90.6*	*29.7*	*179.1*	*6.1*	*-139.9*	*28.4*	*165.1*	*21.0*
o6-13	1,361	-179.1	4.3	52.2	6.1	41.3	8.8	177.4	3.3	78.4	10.2	3.8	12.1
o6-13	*948*	*-178.8*	*4.3*	*52.1*	*6.7*	*41.3*	*8.7*	*177.5*	*3.5*	*78.6*	*10.4*	*3.4*	*12.9*
o6-14	1,015	-179.3	5.8	-123.1	26.6	158.1	22.9	176.3	5.9	73.5	22.4	-157.1	32.7
o6-14	*736*	*-179.6*	*5.6*	*-123.0*	*27.1*	*158.0*	*23.2*	*176.3*	*5.8*	*72.3*	*22.0*	*-156.1*	*33.1*
o6-15	1,079	175.7	6.2	-135.7	21.7	178.7	15.2	-179.0	5.4	-63.2	10.4	-28.5	13.2
o6-15	*716*	*175.9*	*6.5*	*-134.8*	*22.3*	*178.0*	*16.4*	*-178.8*	*4.6*	*-63.6*	*10.6*	*-28.8*	*13.7*
o6-16	1,111	176.9	5.6	-90.8	32.5	155.1	16.4	178.8	4.9	-54.7	6.3	131.9	7.2
o6-16	*775*	*176.9*	*5.5*	*-91.2*	*33.1*	*154.9*	*16.4*	*178.7*	*4.9*	*-55.0*	*6.1*	*131.9*	*7.5*
o6-17	896	179.7	4.0	-65.2	8.7	-35.3	11.4	-178.3	4.0	-81.6	14.0	-39.9	11.1
o6-17	*642*	*179.7*	*3.8*	*-65.7*	*9.2*	*-34.6*	*12.3*	*-178.3*	*3.9*	*-82.7*	*14.5*	*-40.4*	*12.3*
o6-18	864	-179.6	5.5	-89.4	17.8	-10.8	18.9	179.9	5.4	-99.0	32.0	144.3	19.9
o6-18	*629*	*-179.6*	*5.5*	*-89.6*	*18.2*	*-10.8*	*18.8*	*179.9*	*5.4*	*-99.3*	*33.1*	*146.3*	*20.0*
o6-19	813	-179.4	5.2	-91.2	16.6	-7.6	17.8	179.7	5.1	-91.5	26.5	162.9	19.8
o6-19	*574*	*-179.5*	*5.5*	*-90.3*	*16.6*	*-8.3*	*18.0*	*179.8*	*5.2*	*-90.4*	*25.4*	*162.4*	*21.0*
o6-20	969	178.1	7.0	82.9	21.9	-179.8	32.6	-177.0	5.1	-61.4	8.2	-38.3	9.3
o6-20	*689*	*178.3*	*6.4*	*82.1*	*21.5*	*179.6*	*32.9*	*-176.7*	*4.9*	*-61.3*	*8.1*	*-38.1*	*9.3*
o6-21	743	-178.6	4.4	-70.6	12.9	-22.1	13.8	-178.6	5.1	-97.7	16.2	-2.8	18.7
o6-21	*528*	*-178.7*	*4.3*	*-71.0*	*13.2*	*-23.0*	*13.4*	*-178.3*	*5.0*	*-97.8*	*16.0*	*-3.2*	*18.3*
o6-22	766	175.1	5.6	-133.1	14.3	107.8	15.2	-174.8	5.0	60.4	8.1	-123.8	8.2
o6-22	*526*	*175.5*	*5.7*	*-133.0*	*14.9*	*108.3*	*15.3*	*-174.9*	*4.9*	*60.3*	*7.7*	*-123.3*	*8.7*

Table 4.3.8 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+2}		φ_{i+3}		ψ_{i+3}		ω_{i+3}		φ_{i+4}		ψ_{i+4}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o6-1	19,044	-179.9	3.3	-65.9	9.1	-36.0	12.7	179.2	3.7	-77.3	20.3	-30.8	22.2
o6-1	*13,293*	*-179.9*	*3.3*	*-65.9*	*9.2*	*-36.0*	*12.8*	*179.2*	*3.7*	*-77.4*	*20.3*	*-30.7*	*22.3*
o6-2	7,143	-178.6	5.1	-60.4	8.0	-32.4	12.1	-179.9	3.7	-71.8	16.3	-26.2	20.1
o6-2	*4,924*	*-178.8*	*5.0*	*-60.2*	*8.0*	*-32.7*	*12.0*	*-179.9*	*3.6*	*-71.3*	*16.1*	*-26.5*	*19.8*
o6-3	7,070	178.2	4.6	82.3	17.2	16.4	16.2	178.2	4.1	-92.6	20.9	141.0	23.3
o6-3	*4,936*	*178.2*	*4.5*	*82.7*	*17.2*	*16.1*	*16.1*	*178.2*	*4.1*	*-92.5*	*21.1*	*141.2*	*23.3*
o6-4	6,539	-176.6	4.2	-80.4	16.2	-38.9	12.0	-174.5	5.4	-97.5	21.7	-18.7	20.1
o6-4	*4,698*	*-176.7*	*4.2*	*-80.0*	*16.4*	*-39.0*	*12.2*	*-174.7*	*5.5*	*-96.9*	*21.8*	*-18.7*	*20.2*
o6-5	2,742	-178.8	5.4	-136.7	35.5	146.2	34.4	178.4	6.2	-75.5	18.1	132.4	24.7
o6-5	*1,903*	*-178.8*	*5.2*	*-137.1*	*35.0*	*146.3*	*34.4*	*178.6*	*6.6*	*-75.8*	*18.5*	*132.9*	*24.8*
o6-6	2,988	-177.0	5.1	-93.2	18.4	-14.1	14.5	-179.9	5.4	71.0	17.0	20.9	24.4
o6-6	*2,160*	*-177.1*	*5.1*	*-92.6*	*18.6*	*-14.1*	*14.8*	*-179.7*	*5.2*	*71.0*	*16.9*	*21.0*	*24.1*
o6-7	2,291	178.9	4.7	72.4	13.7	23.6	15.7	-179.0	4.4	-89.1	21.1	-12.6	24.4
o6-7	*1,632*	*179.0*	*4.6*	*72.1*	*13.7*	*24.0*	*15.9*	*-178.9*	*4.5*	*-89.8*	*21.4*	*-12.2*	*24.5*
o6-8	2,136	178.1	3.8	-88.9	11.4	2.8	10.6	178.4	4.5	83.3	13.6	8.7	14.8
o6-8	*1,446*	*178.2*	*3.7*	*-88.9*	*11.7*	*3.0*	*10.6*	*178.5*	*4.3*	*83.3*	*13.7*	*8.5*	*15.1*
o6-9	1,877	177.5	3.6	77.6	10.7	5.7	13.6	179.6	4.4	-109.8	19.9	141.8	16.6
o6-9	*1,306*	*177.6*	*3.7*	*77.5*	*10.8*	*5.5*	*14.2*	*179.5*	*4.5*	*-109.5*	*19.8*	*141.7*	*17.2*
o6-10	1,566	-177.7	4.8	-90.2	16.5	-33.9	12.4	-178.3	4.8	-140.0	15.2	85.9	31.7
o6-10	*1,111*	*-177.8*	*4.8*	*-90.4*	*16.3*	*-33.8*	*12.4*	*-178.4*	*4.6*	*-140.4*	*15.2*	*85.2*	*31.6*
o6-11	1,438	-178.0	4.7	-128.6	24.5	105.5	31.4	-177.0	5.8	-63.9	9.8	-24.8	13.2
o6-11	*1,045*	*-178.3*	*4.9*	*-128.4*	*24.5*	*105.9*	*31.4*	*-177.2*	*5.8*	*-63.5*	*9.8*	*-25.0*	*13.4*
o6-12	1,114	-178.9	4.9	-60.0	7.6	-33.9	12.2	-179.4	3.6	-76.2	18.8	-26.5	20.7
o6-12	*777*	*-178.5*	*5.1*	*-59.7*	*8.2*	*-34.3*	*12.1*	*-179.4*	*3.8*	*-75.9*	*18.7*	*-26.8*	*20.8*
o6-13	1,361	179.6	4.0	-113.4	16.8	141.0	13.7	177.4	4.8	-77.5	15.9	131.0	12.5
o6-13	*948*	*179.5*	*4.2*	*-113.1*	*16.9*	*140.7*	*14.0*	*177.5*	*4.8*	*-77.3*	*15.7*	*131.1*	*12.5*
o6-14	1,015	-177.4	5.0	-64.2	12.5	-27.6	16.2	179.8	4.3	-71.5	16.8	-25.4	21.2
o6-14	*736*	*-177.4*	*4.7*	*-64.8*	*12.4*	*-27.1*	*16.2*	*179.8*	*4.3*	*-71.0*	*16.6*	*-25.9*	*20.3*
o6-15	1,079	-176.7	4.4	-106.9	18.0	-9.2	21.2	-177.3	5.5	-143.5	20.8	141.9	30.8
o6-15	*716*	*-176.7*	*4.3*	*-107.0*	*18.0*	*-9.6*	*21.2*	*-177.3*	*5.3*	*-143.6*	*21.0*	*141.1*	*30.6*
o6-16	1,111	179.3	3.4	82.1	12.6	-0.1	14.4	179.6	4.6	-108.7	23.8	150.9	23.0
o6-16	*775*	*179.4*	*3.5*	*82.2*	*12.7*	*-0.3*	*14.5*	*179.7*	*5.4*	*-108.4*	*23.6*	*150.3*	*24.2*
o6-17	896	-174.4	5.2	-101.9	18.6	-15.3	15.0	-177.9	5.7	86.6	17.4	177.3	27.2
o6-17	*642*	*-174.3*	*5.3*	*-102.5*	*17.5*	*-14.1*	*14.8*	*-177.7*	*5.7*	*85.4*	*17.1*	*177.7*	*27.2*
o6-18	864	178.9	4.9	-59.5	9.9	135.1	11.5	178.3	4.0	78.0	16.1	7.4	19.9
o6-18	*629*	*178.9*	*4.8*	*-60.1*	*10.2*	*135.5*	*11.9*	*178.1*	*4.3*	*77.1*	*16.6*	*8.5*	*21.1*
o6-19	813	179.8	5.3	-70.2	13.2	-26.2	19.4	178.5	5.5	-117.2	23.7	123.7	32.2
o6-19	*574*	*179.7*	*5.4*	*-70.1*	*13.3*	*-26.7*	*18.8*	*178.6*	*5.8*	*-116.9*	*24.4*	*122.9*	*32.1*
o6-20	969	-179.7	3.6	-64.4	8.0	-40.0	10.9	178.4	3.6	-70.0	13.6	-36.6	14.3
o6-20	*689*	*-179.6*	*3.8*	*-64.2*	*7.7*	*-40.1*	*10.9*	*178.5*	*3.5*	*-70.4*	*14.0*	*-36.5*	*14.9*
o6-21	743	-178.8	5.2	-113.1	25.7	169.8	15.6	179.1	5.3	-80.4	16.9	-12.6	22.1
o6-21	*528*	*-178.7*	*5.4*	*-112.7*	*25.9*	*169.9*	*17.2*	*179.2*	*5.3*	*-81.2*	*18.0*	*-12.7*	*22.6*
o6-22	766	179.9	3.4	-94.1	11.8	6.3	13.7	-178.0	5.1	-106.3	16.9	139.9	17.6
o6-22	*526*	*179.8*	*3.7*	*-94.4*	*12.5*	*5.8*	*14.4*	*-178.0*	*5.0*	*-105.6*	*17.1*	*138.9*	*17.5*

Table 4.3.8 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+4}		Cα-Cα [Å]	
		∅	±	∅	±
o6-1	19,044	179.9	4.4	6.8	0.7
o6-1	*13,293*	*179.9*	*4.4*	*6.8*	*0.7*
o6-2	7,143	179.4	4.3	6.3	0.9
o6-2	*4,924*	*179.4*	*4.3*	*6.3*	*0.9*
o6-3	7,070	178.5	6.0	6.6	0.7
o6-3	*4,936*	*178.6*	*5.8*	*6.6*	*0.7*
o6-4	6,539	-178.5	5.3	7.3	0.6
o6-4	*4,698*	*-178.6*	*5.2*	*7.3*	*0.6*
o6-5	2,742	179.7	5.5	7.0	0.8
o6-5	*1,903*	*179.7*	*5.5*	*7.0*	*0.8*
o6-6	2,988	179.4	4.8	5.8	1.1
o6-6	*2,160*	*179.4*	*4.6*	*5.8*	*1.1*
o6-7	2,291	179.3	5.5	5.7	0.9
o6-7	*1,632*	*179.5*	*5.4*	*5.8*	*0.9*
o6-8	2,136	178.1	4.6	5.7	0.7
o6-8	*1,446*	*178.1*	*4.6*	*5.7*	*0.7*
o6-9	1,877	177.8	5.1	4.8	0.8
o6-9	*1,306*	*177.9*	*5.0*	*4.8*	*0.8*
o6-10	1,566	-176.5	5.6	7.3	0.6
o6-10	*1,111*	*-176.7*	*5.5*	*7.3*	*0.7*
o6-11	1,438	179.8	4.4	7.5	0.4
o6-11	*1,045*	*179.9*	*4.4*	*7.5*	*0.4*
o6-12	1,114	-179.7	5.6	6.9	0.9
o6-12	*777*	*-179.9*	*4.6*	*6.9*	*0.9*
o6-13	1,361	179.2	5.0	6.9	0.6
o6-13	*948*	*179.4*	*5.1*	*6.9*	*0.6*
o6-14	1,015	178.6	4.1	5.6	1.0
o6-14	*736*	*178.8*	*4.2*	*5.6*	*1.0*
o6-15	1,079	177.3	7.0	5.3	1.1
o6-15	*716*	*177.5*	*6.8*	*5.3*	*1.1*
o6-16	1,111	178.4	6.4	6.8	0.9
o6-16	*775*	*178.5*	*6.2*	*6.8*	*0.9*
o6-17	896	-177.9	5.2	6.1	1.0
o6-17	*642*	*-177.6*	*4.9*	*6.1*	*1.0*
o6-18	864	179.4	5.3	6.3	0.9
o6-18	*629*	*179.6*	*5.1*	*6.3*	*0.9*
o6-19	813	-179.8	5.6	6.9	0.9
o6-19	*574*	*-179.5*	*6.0*	*6.9*	*0.9*
o6-20	969	179.3	4.0	7.1	0.7
o6-20	*689*	*179.3*	*4.0*	*7.1*	*0.7*
o6-21	743	-179.5	5.0	7.0	0.8
o6-21	*528*	*-179.7*	*5.2*	*7.0*	*0.8*
o6-22	766	176.0	5.3	4.6	0.4
o6-22	*526*	*176.1*	*5.4*	*4.6*	*0.4*

Reverse Turns

Reverse turns are defined by a hydrogen bond between the backbone nitrogen atom of residue *i* and the backbone oxygen atom of residue *i+x*. They comprise two to six residues and their functional analysis is restricted to a few studies underpinning their meaning for protein folding.[429] In the following, the reverse turn types and their underlying torsion angles and residue preferences are discussed. Compared to the normal and open turn classes, they have no impact on the assignment of helices with SCOT. However, some reverse turn categories are used for the assignment of the β-strand boundaries.

Reverse 2 Turns

The U-Matrix of the reverse 2 turns (Figure 4.3.14) shows a clear differentiation between twelve classes although the number of representative turns is considerably lower as compared to the initial turn classification study with Secbase. The reason is the more restrictive hydrogen bond criterion used in this study.

Nonetheless, correspondences to the formerly defined turn types were found. The turn type r2-1 corresponds to types I and II, r2-2 to type III, and r2-5 to type IV of the former assignments. Intriguingly, the turns of type r2-4, r2-11, and r2-12 were originally classified as broken structures. For type r2-10, an atypical dihedral angle of the peptide bond was found. Broken turns could not be found in the training dataset for the new ESOMs as they arose from turn structures with an incomplete backbone in the Secbase-derived assignments. In general, the new classification is better suited to assign distinct sets of reverse 2 turn types. The previously assigned types do not cluster in distinct valleys of the U-Matrix and the types r2-3 and r2-6 to r2-10 are new types which remained unassigned with the Secbase-methodology (no type). However, only types r2-1 to r2-4 are significantly represented in the dataset and the classification of other types should be verified using an even larger dataset of structures.

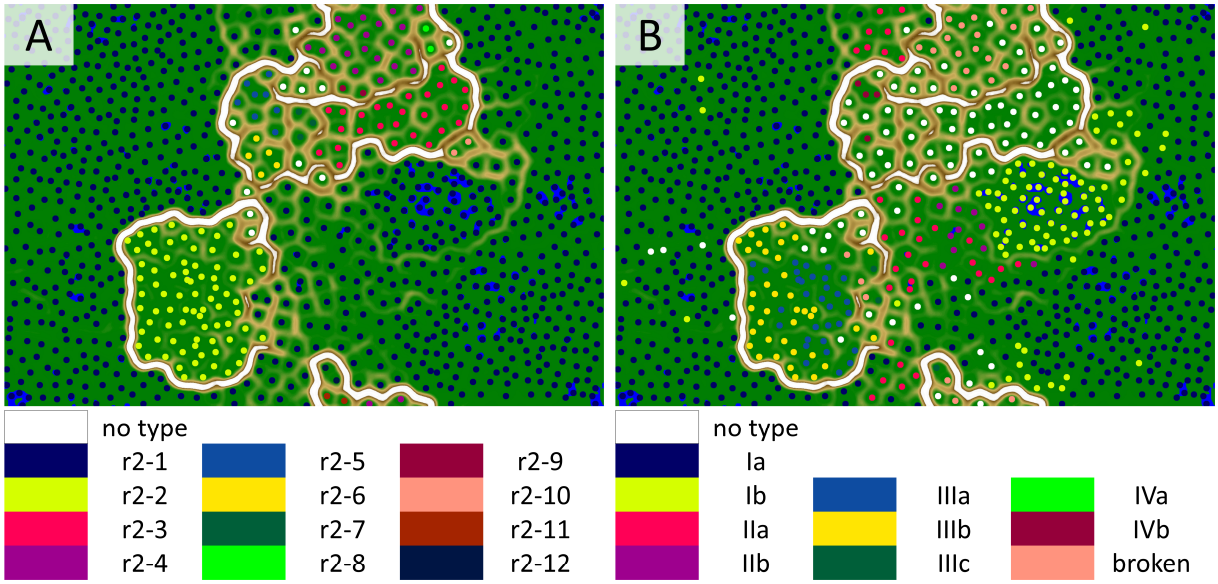


Figure 4.3.14. Reverse 2 turn clustering. The U-Matrix for reverse 2 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown.

The mean dihedral angles and standard deviations for the distinct types are given in Table 4.3.9. The types r2-3 to r2-12 are characterised by a *cis* peptide bond. The results for the training and test set are in good agreement for the largest classes r2-1, r2-2, r2-3, and r2-4. The hydrogen bond energies are lower than those for most other hydrogen-bonded turn structures. The most stable interactions are found for types r2-1 and r2-4. For the turn type r2-1, a clear overrepresentation of aromatic residues (Trp and Tyr) was observed at positions *i* and *i+1* (Table A3.14). Stacking interactions between these residues support the weak hydrogen bond interaction and lead to the stable conformation of the turn type. Gly and Pro residues are over-represented in the types r2-2/r2-4/r2-5 and r2-3, respectively, leading to the atypical dihedral angles usually found in the coiled regions of protein structures. The low hydrogen bond energies for all reverse 2 turns can be attributed to conformational restrictions of the protein backbone which prevent an optimum hydrogen bond geometry.

Table 4.3.9. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 2 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel. The asterisk highlights turn types which could not be found in the test set.

type	|turns|	φ_i		ψ_i		ω_i		φ_{i+1}		ψ_{i+1}		E_{hb} [kcal/mol]	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r2-1	810	-147.4	16.5	-77.3	14.4	162.2	9.6	-82.8	14.7	175.7	12.3	-2.1	1.0
r2-1	*26*	*-157.1*	*22.4*	*-68.5*	*17.2*	*162.6*	*7.8*	*-85.4*	*15.1*	*175.6*	*10.6*	*-2.1*	*1.1*
r2-2	79	-155.9	16.6	56.9	18.3	-165.2	7.3	82.5	15.4	-178.8	13.8	-1.5	0.8
r2-2	*9*	*-156.3*	*9.8*	*60.9*	*13.5*	*-165.5*	*8.0*	*80.9*	*11.1*	*179.0*	*12.7*	*-1.6*	*0.8*
r2-3	27	-111.6	27.9	-16.0	28.1	7.2	9.2	-81.8	16.7	-41.7	22.7	-1.4	0.7
r2-3	*3*	*-112.9*	*12.5*	*-15.1*	*20.9*	*6.4*	*6.1*	*-82.4*	*17.1*	*-52.0*	*21.3*	*-1.2*	*0.6*
r2-4	23	-129.9	28.8	-61.5	17.4	3.8	15.1	76.8	24.4	95.7	27.9	-2.2	1.6
r2-4	*1*	*-127.3*	*n/a*	*-67.1*	*n/a*	*-1.2*	*n/a*	*73.6*	*n/a*	*84.0*	*n/a*	*-2.4*	*n/a*
r2-5	8	122.8	10.6	92.5	15.9	1.5	6.1	-69.1	18.8	166.1	29.9	-1.1	0.5
r2-5	*1*	*126.1*	*n/a*	*113.6*	*n/a*	*-8.7*	*n/a*	*-51.2*	*n/a*	*161.0*	*n/a*	*-0.9*	*n/a*
r2-6	5	-159.4	14.2	43.3	21.6	-1.7	7.4	-82.0	7.3	-177.0	11.0	-0.9	0.3
r2-6	*1*	*-141.5*	*n/a*	*31.7*	*n/a*	*-0.8*	*n/a*	*-103.8*	*n/a*	*-174.6*	*n/a*	*-1.4*	*n/a*
r2-7*	3	104.3	23.2	76.0	6.1	-0.6	9.6	-96.0	14.7	-59.7	12.8	-1.1	0.8
r2-8*	2	-98.6	8.3	-75.1	4.0	6.3	9.0	21.1	19.8	-122.1	6.5	-1.3	0.5
r2-9*	2	-165.5	27.4	-34.6	0.7	14.8	2.6	70.0	5.3	23.7	3.8	-0.9	0.3
r2-10*	2	-135.5	9.4	-61.0	3.2	128.3	9.3	-119.6	9.8	-86.6	2.6	-1.6	0.8
r2-11*	2	-163.4	10.8	-70.8	0.9	1.4	0.6	76.9	0.5	177.2	17.1	-1.2	0.5
r2-12*	2	154.3	6.4	-30.2	6.8	-21.8	3.9	120.6	0.9	178.5	3.2	-0.9	0.3

Reverse 3 Turns

Figure 4.3.15 shows the U-Matrix for the classification of reverse 3 turns. In contrast to the Secbase-derived types, six types are distinguished. Two of them (r3-5 and r3-6) are characterised by a *cis* peptide bond between residue positions *i* and *i+1* which was previously not observed (turns of these types have no type assignment using the former classification). The turn type r3-5 shows a weaker hydrogen bond and Pro residues are overrepresented at position *i+1* (see Table A3.15 in Appendix 3). However, the hydrogen bond energies are significantly higher as compared to those of the reverse 2 turns (see Table 4.3.10).

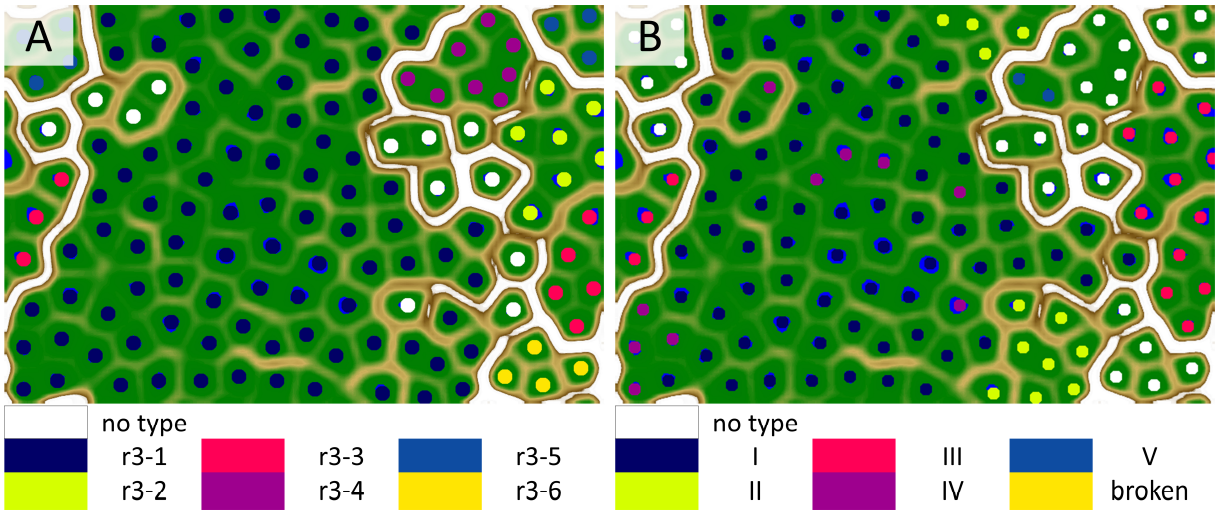


Figure 4.3.15. Reverse 3 turn clustering. The U-Matrix for reverse 3 turns with the projected best matches of the training site coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown.

Table 4.3.10. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 3 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	φ_i		ψ_i		ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r3-1	88	-144.0	29.1	-116.7	24.0	173.8	4.0	-66.6	12.4	-31.8	11.2	173.9	6.0
r3-1	*57*	*-144.9*	*30.0*	*-116.1*	*23.2*	*173.7*	*5.0*	*-64.6*	*10.4*	*-33.9*	*11.4*	*173.6*	*6.7*
r3-2	7	-127.8	9.7	31.7	8.5	-169.2	4.0	111.3	7.8	-24.9	10.9	172.4	8.1
r3-2	*4*	*-125.4*	*7.6*	*32.2*	*11.1*	*-168.9*	*1.0*	*99.8*	*9.8*	*-11.2*	*12.2*	*168.7*	*6.3*
r3-3	8	-145.9	13.0	87.5	9.8	-172.5	2.2	73.8	4.7	-57.8	12.5	173.6	3.4
r3-3	*6*	*-145.5*	*12.7*	*89.8*	*6.1*	*-166.7*	*5.1*	*68.1*	*6.0*	*-52.7*	*23.8*	*170.7*	*4.8*
r3-4	8	-160.0	4.0	97.5	10.9	-173.1	5.4	49.8	3.9	51.2	5.6	-176.1	2.5
r3-4	*6*	*-152.3*	*10.1*	*99.4*	*9.4*	*-175.0*	*3.4*	*50.7*	*3.7*	*50.4*	*4.7*	*-177.1*	*1.7*
r3-5	7	-120.1	22.8	23.0	25.6	6.6	7.5	-86.0	8.7	-8.6	8.0	176.5	3.1
r3-5	*2*	*-101.5*	*15.7*	*-4.9*	*3.5*	*15.6*	*8.1*	*-83.1*	*11.7*	*-6.6*	*6.2*	*176.2*	*2.7*
r3-6	4	81.1	12.2	109.2	16.8	10.7	10.3	-78.9	9.6	-19.7	10.0	179.9	4.8
r3-6	*3*	*85.5*	*9.5*	*117.4*	*13.3*	*7.3*	*9.9*	*-77.1*	*9.7*	*-20.5*	*5.6*	*178.7*	*0.9*

type	|turns|	φ_{i+2}		ψ_{i+2}		E_{hb} [kcal/mol]	
		∅	±	∅	±	∅	∅
r3-1	88	-126.2	15.2	166.6	12.0	-4.5	1.6
r3-1	*57*	*-124.8*	*16.6*	*167.2*	*14.7*	*-4.4*	*1.7*
r3-2	7	-139.0	15.2	162.0	8.8	-3.4	1.1
r3-2	*4*	*-150.5*	*16.8*	*163.3*	*7.4*	*-4.5*	*0.5*
r3-3	8	-114.8	23.8	165.4	6.1	-4.1	1.7
r3-3	*6*	*-114.9*	*31.5*	*165.6*	*4.0*	*-4.3*	*1.3*
r3-4	8	120.0	11.7	168.9	16.9	-4.7	1.1
r3-4	*6*	*114.4*	*19.2*	*167.9*	*16.2*	*-4.4*	*1.8*
r3-5	7	-77.6	19.8	153.4	20.9	-2.4	0.9
r3-5	*2*	*-57.4*	*0.5*	*137.4*	*1.6*	*-2.9*	*0.6*
r3-6	4	-85.1	16.0	151.0	17.8	-4.0	1.8
r3-6	*3*	*-89.9*	*9.7*	*154.1*	*20.8*	*-4.3*	*2.1*

The number of detected reverse 3 turns is low and Gly and Pro residues are overrepresented in nearly all classes (see Table A3.15 in Appendix 3). For this turn category, the following correspondences to the original turn types were identified. The type r3-1 includes the previously found types I, II, and IV. This agrees with the dihedral angle ranges of these classes. The turn type III was subdivided into the clusters r3-2 and r3-3 which differ significantly with respect to the angles φ_i, ψ_{i+1}, and φ_{i+1}. The turn type r3-4 corresponds to type V.

Reverse 4 Turns

The number of reverse 4 turns is higher than that for the previously discussed reverse turns. As the conformational degree of freedom increases for this turn length, the number of turn types rises. The U-Matrix for the final classification is depicted in Figure 4.3.16. Only 11 types are classified in contrast to the Secbase-derived 13 turn types for this category. A closer look at the distribution of the initial and the new turn types across the U-Matrix gives a plausible reason. The reverse 4 turns of types I', XII, and XIII are now assigned to the cluster r4-1. Similarly, the original turn types IIIa/b, IVa/b, Va/b, and Xa/b are classified as one turn type (r4-2). These fusions can be explained by the dihedral angles of residues *i* and *i+3* which are approximately 180° (Table 4.3.11). The introduced angle transformation for clustering reflects the small distances in the range of 180°/-180°. The novel type r4-3 comprises the former types II' and XI. Additionally, corresponding types were identified: r4-4 and VII, r4-5 and IX, r4-6 and VIII, and r4-8 and VI (with a *cis* peptide bond between residue *i+1* and *i+2*).

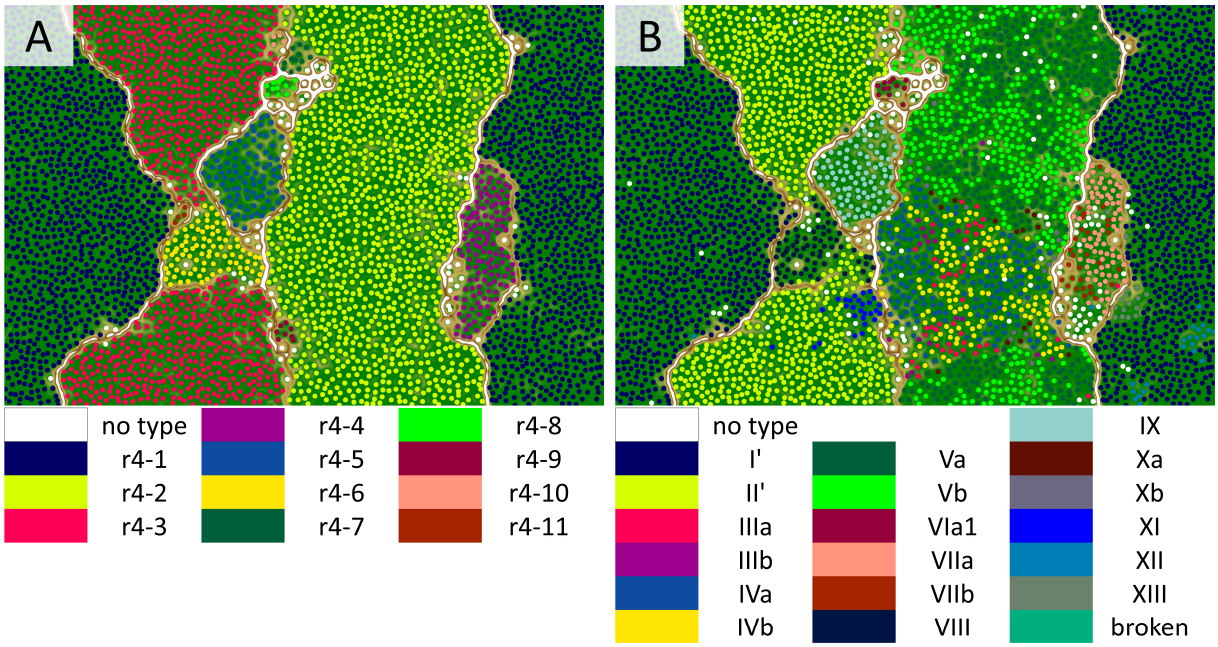


Figure 4.3.16. Reverse 4 turn clustering. The U-Matrix for reverse 4 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown.

The clustering of the reverse 4 turns illustrates the benefits of using two transformed dihedral angles for an ESOM-based clustering of angular data. The turn type VII was originally found in two separate valleys of the underlying U-Matrix (leading to the types VIIa and VIIb) as the ψ_i angles range between -150° and -180° or 180° and 160°. Their numerical distance is 310°, while their distance in angular space is only 50°. Both types are now grouped in one common cluster.

The corresponding means and standard deviations of the angular data for the new type r4-4 (Table 4.3.11) support the finding that both types can be assigned to one class. Reverse 4 turns are stable (see hydrogen bond energies in Table 4.3.11) and although Gly residues are still overrepresented in most turn types, further residues are prevalent in the classified reverse 4 turns (see Table A3.16 in Appendix 3).

The Ramachandran plots of the highest occupied turn types show that types r4-1, r4-2, r4-3, and r4-4 are subsets of the normal 4 types n4-3, n4-1, n4-4, and n4-2, respectively (Figure 4.3.17). Most of these four residue turns are characterised by an additional reverse hydrogen bond between residue *i* and *i+3*. They are often observed as connecting elements between adjacent β-strands which is in line with the dihedral angle ranges of the residues at the positions *i* and *i+3* which are situated in the immediate proximity of regions corresponding to the extended β-strand conformation. They stabilise the kinked regions connecting hydrogen-bonded β-strands. This explains the high number of structures of these turn types in the dataset. The frequent occurrence of Gly and Pro residues (see Table A3.16) at varying positions enables the reversal of the direction of the polypeptide chain.

Table 4.3.11. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 4 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	φ_i		ψ_i		ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r4-1	1,654	-135.5	16.3	123.4	12.7	-179.3	4.3	51.9	5.7	41.6	7.8	177.4	3.3
r4-1	*1,150*	*-135.0*	*16.3*	*123.2*	*12.6*	*-179.1*	*4.4*	*51.8*	*5.6*	*41.7*	*7.8*	*177.4*	*3.2*
r4-2	1,208	-130.1	21.5	-178.1	11.6	-179.5	4.4	-61.5	8.2	-29.1	12.2	-177.3	4.1
r4-2	*818*	*-128.9*	*22.1*	*-178.5*	*12.4*	*-179.3*	*4.5*	*-62.1*	*8.5*	*-28.7*	*12.1*	*-177.4*	*4.3*
r4-3	828	-132.6	14.9	107.7	16.1	-174.8	5.1	60.6	7.6	-122.9	10.2	179.9	3.5
r4-3	*551*	*-133.1*	*15.0*	*107.7*	*17.2*	*-174.9*	*5.0*	*60.4*	*7.6*	*-122.0*	*11.2*	*179.9*	*3.7*
r4-4	137	-141.4	23.7	174.0	12.2	178.2	4.7	-58.5	8.9	120.0	9.6	179.1	3.2
r4-4	*99*	*-140.3*	*21.9*	*173.7*	*12.3*	*178.9*	*4.3*	*-59.2*	*9.5*	*120.3*	*10.7*	*179.4*	*2.9*
r4-5	94	85.5	13.7	-6.8	15.2	178.7	5.1	-115.0	13.0	-60.9	13.5	178.4	3.7
r4-5	*68*	*85.7*	*12.8*	*-7.5*	*14.6*	*178.8*	*4.7*	*-113.1*	*14.3*	*-61.2*	*12.6*	*178.7*	*4.6*
r4-6	89	-104.2	11.2	10.6	10.1	-178.1	4.4	98.6	14.5	-36.2	26.1	179.1	2.9
r4-6	*61*	*-103.4*	*11.8*	*10.8*	*9.5*	*-177.4*	*4.3*	*98.2*	*13.1*	*-32.8*	*25.3*	*178.9*	*4.4*
r4-7	12	-102.7	18.5	-86.5	22.4	176.3	7.4	-105.4	13.6	-87.7	16.8	171.3	8.1
r4-7	*5*	*-104.8*	*14.9*	*-78.2*	*19.5*	*175.8*	*4.0*	*-112.2*	*10.8*	*-92.6*	*27.3*	*179.4*	*1.3*
r4-8	12	-136.1	14.1	88.5	14.6	-172.2	7.1	-44.5	8.6	134.9	5.8	4.3	6.4
r4-8	*5*	*-141.6*	*8.3*	*94.1*	*12.6*	*-172.7*	*5.8*	*-45.5*	*5.4*	*136.2*	*4.0*	*3.7*	*3.5*
r4-9	8	-107.8	13.2	-177.3	6.8	176.6	6.1	-61.0	5.3	-39.7	10.5	173.8	5.3
r4-9	*5*	*-107.2*	*9.2*	*175.6*	*9.4*	*179.0*	*1.7*	*-59.1*	*3.7*	*-41.9*	*9.3*	*172.2*	*3.7*
r4-10	4	114.9	12.3	167.3	15.9	175.8	2.3	62.1	7.7	-122.2	4.0	-179.7	3.6
r4-10	*5*	*117.4*	*12.9*	*165.2*	*16.3*	*177.5*	*3.2*	*60.8*	*5.4*	*-120.8*	*5.5*	*-179.7*	*2.3*
r4-11	4	-105.2	2.6	13.4	9.7	-179.6	4.3	78.3	11.3	94.8	23.8	-177.6	2.2
r4-11	*1*	*-93.6*	*0.0*	*10.7*	*0.0*	*-164.6*	*0.0*	*81.9*	*0.0*	*119.2*	*0.0*	*176.3*	*0.0*

Table 4.3.11 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 4 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	φ_{i+2}		ψ_{i+2}		ω_{i+2}		φ_{i+3}		ψ_{i+3}		E_{hb} [kcal/mol]	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r4-1	1,654	78.4	9.9	3.5	12.5	179.6	4.1	-114.5	16.3	141.0	15.0	-4.6	1.1
r4-1	*1,150*	*78.4*	*9.9*	*3.5*	*12.7*	*179.6*	*4.1*	*-114.5*	*16.3*	*140.8*	*15.6*	*-4.6*	*1.1*
r4-2	1,208	-105.7	17.7	-6.9	21.6	-178.0	5.3	-165.4	39.7	157.7	36.1	-4.2	1.9
r4-2	*818*	*-106.3*	*17.1*	*-6.4*	*21.1*	*-177.9*	*5.2*	*-165.8*	*40.1*	*157.7*	*36.7*	*-4.0*	*1.9*
r4-3	828	-94.6	12.7	6.1	14.5	-178.0	5.1	-107.9	18.1	140.0	20.3	-5.1	1.3
r4-3	*551*	*-94.6*	*13.1*	*5.1*	*14.3*	*-178.1*	*6.9*	*-106.7*	*18.6*	*138.8*	*19.4*	*-5.1*	*1.3*
r4-4	137	74.9	14.4	9.1	16.3	-178.6	4.2	-150.0	20.4	133.2	37.7	-3.3	1.8
r4-4	*99*	*75.1*	*14.2*	*9.4*	*16.5*	*-179.2*	*3.9*	*-152.2*	*22.8*	*133.0*	*40.4*	*-3.3*	*1.8*
r4-5	94	-96.7	17.8	-24.7	17.0	179.0	5.5	-138.6	20.0	147.2	20.6	-4.6	1.5
r4-5	*68*	*-97.4*	*17.2*	*-24.1*	*16.5*	*179.5*	*4.9*	*-142.2*	*18.9*	*148.7*	*19.0*	*-4.6*	*1.6*
r4-6	89	-110.1	23.2	-30.1	17.3	-179.3	3.8	-139.6	17.0	155.8	15.8	-4.5	1.4
r4-6	*61*	*-112.4*	*21.5*	*-32.5*	*16.5*	*-178.8*	*3.9*	*-138.9*	*18.7*	*158.3*	*14.1*	*-4.5*	*1.4*
r4-7	12	-100.5	19.6	1.3	17.8	-178.5	2.3	-105.4	19.7	139.2	17.2	-4.0	1.4
r4-7	*5*	*-97.1*	*25.8*	*-9.5*	*12.8*	*-179.3*	*1.2*	*-116.3*	*32.3*	*144.3*	*15.9*	*-4.1*	*2.2*
r4-8	12	-89.3	4.8	16.9	14.2	-176.8	5.6	-104.3	18.2	149.0	12.3	-4.3	1.5
r4-8	*5*	*-90.3*	*4.5*	*17.8*	*9.2*	*-177.3*	*3.8*	*-106.5*	*18.5*	*153.1*	*5.6*	*-4.9*	*0.2*
r4-9	8	-131.2	16.2	120.2	12.6	179.0	3.0	77.3	12.9	-171.5	19.4	-3.9	1.3
r4-9	*5*	*-127.0*	*12.6*	*121.9*	*16.8*	*-179.8*	*4.5*	*83.8*	*14.1*	*178.0*	*28.9*	*-4.4*	*1.2*
r4-10	4	-71.5	7.0	-13.2	8.4	178.8	2.2	-114.0	6.7	94.0	18.6	-3.6	2.2
r4-10	*5*	*-75.9*	*7.8*	*-11.5*	*6.5*	*-178.6*	*2.1*	*-110.6*	*11.2*	*91.7*	*13.5*	*-4.3*	*1.9*
r4-11	4	92.5	28.8	-6.2	18.6	176.3	3.5	-154.1	16.8	140.0	7.3	-5.9	1.5
r4-11	*1*	*54.7*	*0.0*	*-14.6*	*0.0*	*172.9*	*0.0*	*-120.9*	*0.0*	*125.3*	*0.0*	*-3.3*	*0.0*

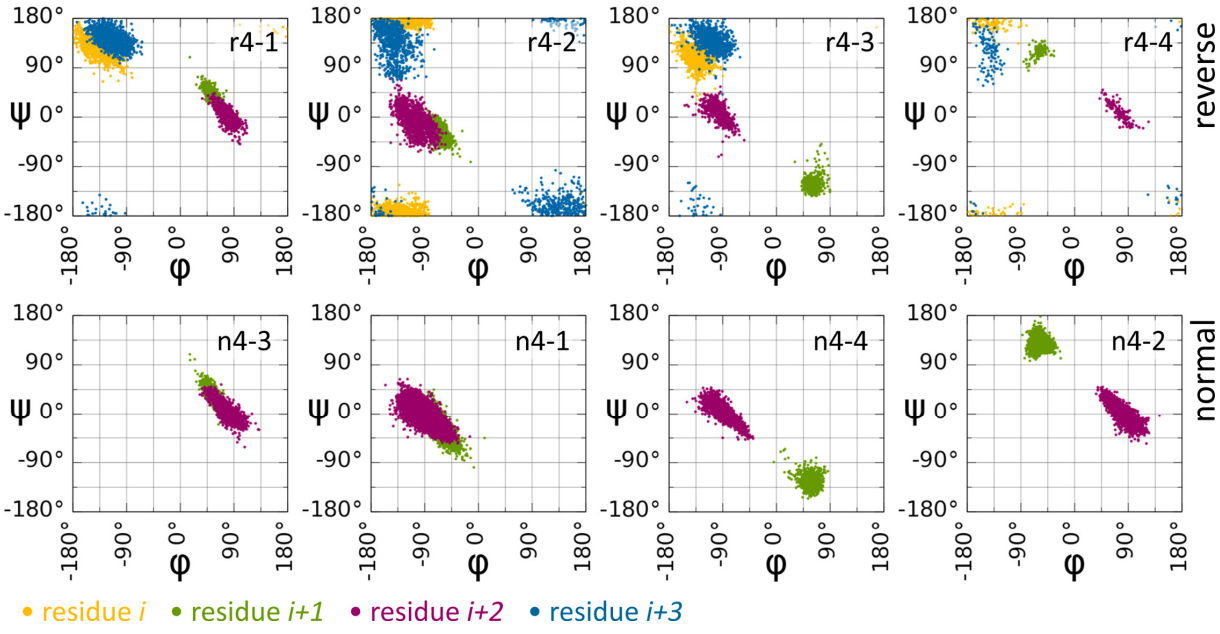


Figure 4.3.17. Ramachandran plots of the highest occupied normal and reverse 4 turns within the training set. The relationships between the reverse 4 and normal 4 turn types (correspondences between ψ_{i+1}, φ_{i+1}, ψ_{i+2}, and φ_{i+2}) can be analysed based on their distributions of the ψ and φ angles.

An analysis of the locations of the reverse turns in protein structures underlines their unique impact on β-sheet formation (see Table 4.3.12). All turns are primarily found at the N- and C-terminus of β-strands. As the strand assignment by SCOT relies on regular continuous hydrogen bonding patterns of the main-chain, four reverse turn categories can be used to avoid the assignment of long strands with high bending angles.

Table 4.3.12. The relationship between reverse turns and β-strands. The number of reverse turns that overlap with DSSP-assigned β-strands by one, two, or three residues in the training set are shown.

Turn Category (Number of Turns)	Reverse 3 (134)	Reverse 4 (4,113)	Reverse 5 (2,892)	Reverse 6 (1,869)
N-terminal overlap (1/2/3)	3	3,185/4	757/1	915/7/1
C-terminal overlap (1/2/3)	1	2,881	653/222	989/25/3

Reverse 5 Turns

A comparison between the novel turn types and the Secbase-derived type assignments becomes difficult for the reverse 5 turns. The U-Matrix with the corresponding best matches and their initial and new type assignments (Figure 4.3.18) shows that some types were combined (e.g., Ia to Ic, IIa and IIb, etc.), while other types were divided into several clusters (e.g., type V, type VII). Some new turn types were not assigned by Secbase (e.g., type r5-14). The conformations a and b of the previously assigned turn types could be found in one cluster due to the angle transformations applied for ESOM training. Nonetheless, more types are classified, and the map enables a clearer differentiation between distinct clusters due to the higher number of neurons used for ESOM training (3 times the number of turns in the former study and 100 times the number of turns in this study).

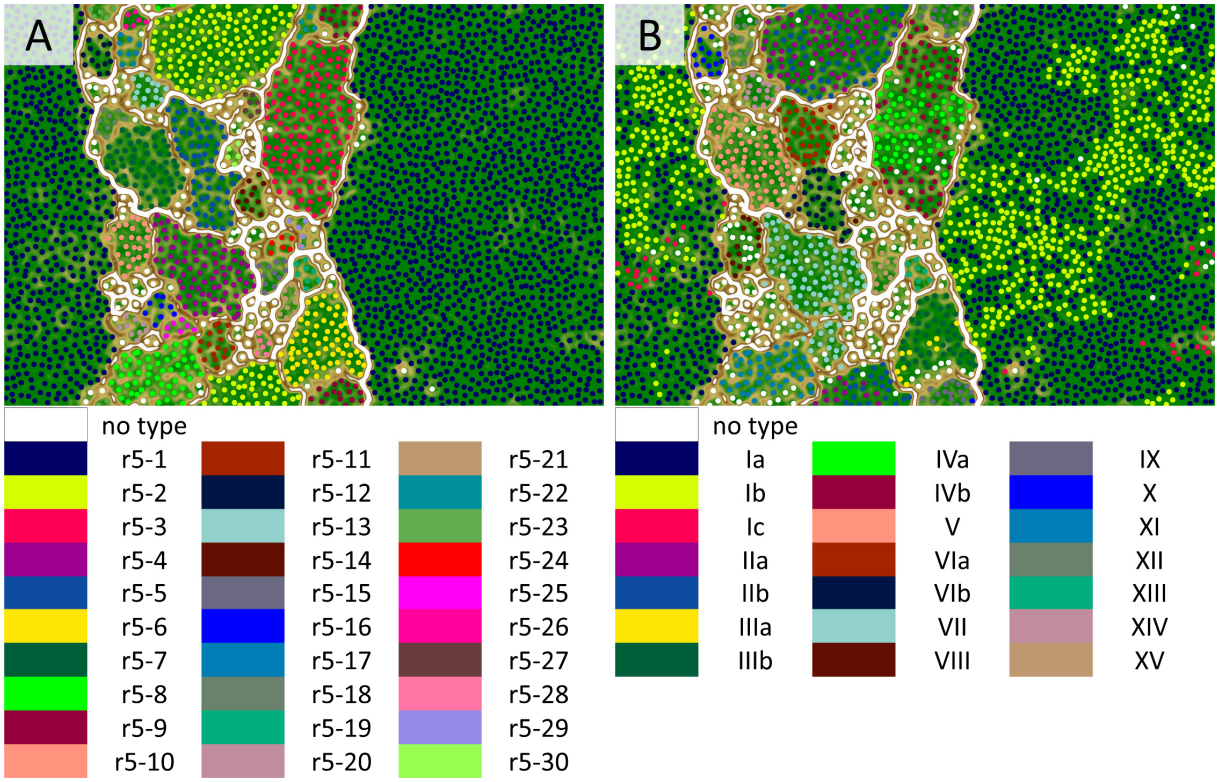


Figure 4.3.18. Reverse 5 turn clustering. The U-Matrix for reverse 5 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown.

Despite these differences, the following turn type correspondences (for the largest clusters) were derived: r5-1 to Ia-c, r5-2 to IIa/IIb, r5-3 to IVa/IVb, r5-4 to VII, r5-5 to VIa/VIb, r5-6 to IIIa/b, and r5-7 to V. Given the results presented in Figure 4.3.18 and Table 4.3.13, the new classification of turn structures assigns new types and merges multiple of the Secbase-assigned types with similar dihedral angles.

Table 4.3.13 summarises the means and standard deviations of the dihedral angles and hydrogen bond energies for the 22 highest occupied of the 30 classified turn types. The characteristics of the types r5-23 to r5-30 can be found in Table A3.17.

The high number of small clusters points towards the potential dataset dependence of the classification. Nevertheless, all types show distinct angular ranges and similar mean dihedral angles for the turns of the training and the test set. The standard deviations of the dihedral angles are slightly higher than those of the previously discussed turn types. The types r5-4, r5-16, and r5-21 are characterised by high standard deviations. They are in line with the low energy of their hydrogen bond which usually restricts the conformational space of hydrogen-bonded turns and the higher degree of freedom for the backbone of turns with a length of five residues. The turn types r5-15, r5-24 (they correspond to the Secbase-assigned type XV), and r5-29 harbour a *cis* peptide bond between residues *i+1* and *i+2*. The type r3-30 is characterised by a *cis* peptide bond between residues *i+2* and *i+3* (see also Table A3.17).

Reverse 5 turns are less frequently found in the training set as compared to the four residue turns. Only a handful of investigations of the impact of this turn category on protein structure exist. However, reverse turns are commonly found interconnecting two SSEs.[430]

In particular, the type r5-1 is found at the N- and C-termini of strands and stabilises the conformation of the residue fragment connecting two parallel or antiparallel strands (see also Table 4.3.12). This turn type encompasses ten times more turns than the second highest occupied type. The turns do not span other reverse turns, but r5-1 turns commonly co-occur with normal turns of type n4-1. In such cases, both turns start at the same residue and stabilise the sharp bend between two adjacent strands. The less frequent type r5-2 usually spans n4-1 turns, but both turns share a common C-terminal residue. The types r5-1 and r5-2 are characterised by a high prevalence for the residues Asn, Asp, Ser, and Thr (see Table A3.18 in Appendix 3). Gly residues are preferentially found at position *i+3* of r5-1 turns, while Pro residues are significantly overrepresented at position *i+2* of r5-2 turns. Both residues enable the extremely bent character of these turn types and highlight their unique character.

For the remaining types, similar co-occurrences with regular SSEs are observed and the types r5-3 to r5-8 are characterised by overrepresented Pro and Gly residues at varying positions. The turn types r5-4 and r5-5 are again subclasses of the corresponding normal 5 types n5-1 and n5-2 (see Figure A3.1). They are commonly found at identical positions connecting two adjacent strands. This stabilisation of the reverse turns by additional O_i to NH_{i+x} hydrogen bonds of the normal turns partially leads to the consistent geometries for the highest occupied types of this turn category with a length of five residues.

Table 4.3.13. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 5 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	φ_i		ψ_i		ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r5-1	1,902	-87.0	14.8	173.4	11.2	-178.3	4.3	-62.6	7.5	-22.4	10.8	177.9	3.8
r5-1	*1,272*	*-86.6*	*14.2*	*173.1*	*11.2*	*-178.3*	*4.2*	*-62.4*	*7.4*	*-22.6*	*10.9*	*178.0*	*3.7*
r5-2	170	-96.0	14.7	-33.9	17.3	176.0	5.4	-156.2	12.5	-178.0	8.4	-178.4	4.9
r5-2	*101*	*-97.9*	*14.8*	*-31.7*	*17.3*	*176.1*	*5.3*	*-157.0*	*12.3*	*-178.4*	*8.7*	*-178.3*	*5.0*
r5-3	138	-111.6	20.0	-177.1	21.9	179.1	5.5	-59.8	11.6	147.6	13.4	178.4	5.1
r5-3	*80*	*-113.7*	*20.0*	*-173.4*	*21.8*	*179.3*	*5.3*	*-59.6*	*12.2*	*147.4*	*13.2*	*178.5*	*4.8*
r5-4	82	-117.2	27.0	146.1	16.0	-178.0	4.3	-64.0	8.6	-30.8	13.0	-178.2	4.3
r5-4	*64*	*-119.2*	*26.3*	*146.7*	*14.9*	*-178.7*	*4.5*	*-62.5*	*8.5*	*-32.5*	*12.9*	*-177.8*	*4.3*
r5-5	69	-119.4	18.8	116.3	14.3	-177.3	5.2	-51.8	7.0	129.2	7.0	179.4	2.6
r5-5	*47*	*-115.9*	*19.5*	*116.3*	*14.5*	*-177.5*	*5.1*	*-52.5*	*6.0*	*130.4*	*6.3*	*179.4*	*2.6*
r5-6	67	-94.9	16.3	-29.1	17.9	176.6	5.2	-162.7	17.9	171.0	8.7	179.7	4.8
r5-6	*50*	*-94.8*	*16.3*	*-31.7*	*21.7*	*176.7*	*4.5*	*-161.7*	*18.9*	*170.9*	*9.0*	*-179.7*	*3.6*
r5-7	61	-101.2	20.1	123.7	10.9	178.9	4.2	54.2	6.1	28.3	9.1	-179.5	3.6
r5-7	*34*	*-102.8*	*19.5*	*122.2*	*12.7*	*178.9*	*4.5*	*53.6*	*5.2*	*30.2*	*6.4*	*-179.1*	*3.5*
r5-8	56	-124.5	14.7	111.0	22.0	-175.5	5.9	60.5	11.3	-148.9	18.8	-179.5	5.0
r5-8	*32*	*-121.9*	*14.5*	*108.4*	*23.5*	*-175.5*	*5.5*	*61.7*	*11.5*	*-149.8*	*21.1*	*-179.1*	*5.1*
r5-9	29	-102.6	11.9	1.9	13.0	-179.2	5.9	95.5	14.1	156.7	13.4	178.2	4.4
r5-9	*19*	*-101.1*	*10.3*	*-1.6*	*11.5*	*-177.7*	*4.8*	*96.5*	*13.4*	*156.8*	*12.4*	*178.8*	*2.7*
r5-10	26	-93.8	16.6	142.3	13.2	-178.6	6.3	-59.6	5.6	-31.9	9.8	-179.6	3.0
r5-10	*19*	*-90.1*	*17.7*	*142.9*	*11.4*	*-179.7*	*5.1*	*-59.8*	*5.1*	*-32.2*	*9.3*	*-179.5*	*4.1*
r5-11	17	-111.6	16.5	49.1	17.3	177.5	5.7	56.2	7.4	-129.7	6.2	-178.9	3.7
r5-11	*13*	*-113.4*	*18.7*	*53.9*	*17.4*	*177.6*	*4.9*	*54.9*	*7.0*	*-128.2*	*5.3*	*-178.4*	*3.0*
r5-12	15	-91.1	10.1	-5.6	7.3	178.7	4.4	83.2	11.9	3.3	8.8	-178.2	3.1
r5-12	*6*	*-90.6*	*7.4*	*-1.2*	*1.3*	*-180.0*	*3.7*	*76.5*	*9.7*	*10.1*	*10.1*	*-179.6*	*2.1*
r5-13	16	-94.3	12.5	-55.4	13.3	176.3	3.9	-164.8	7.0	141.6	13.1	179.2	3.2
r5-13	*12*	*-96.8*	*12.8*	*-54.4*	*11.1*	*177.6*	*4.4*	*-166.9*	*7.6*	*143.0*	*14.5*	*177.8*	*3.5*
r5-14	12	-110.3	42.9	138.3	21.7	-179.6	2.5	-59.1	6.6	127.1	7.5	-176.9	3.9
r5-14	*7*	*-117.5*	*36.9*	*137.5*	*15.5*	*178.9*	*2.3*	*-61.8*	*6.0*	*130.3*	*7.6*	*-175.6*	*3.2*
r5-15	11	-139.3	16.7	89.0	18.0	-180.0	5.1	-70.6	4.7	141.2	6.5	2.6	5.9
r5-15	*9*	*-141.4*	*19.4*	*93.4*	*19.5*	*-178.4*	*4.0*	*-72.6*	*4.1*	*141.7*	*8.9*	*-1.2*	*3.7*
r5-16	9	-138.1	24.6	137.6	20.7	-173.4	5.0	-62.0	9.3	-28.5	8.3	-178.0	3.8
r5-16	*5*	*-147.6*	*24.1*	*137.7*	*17.1*	*-177.0*	*3.4*	*-60.2*	*8.8*	*-27.2*	*3.5*	*-177.9*	*1.4*
r5-17	9	-106.8	12.7	1.8	6.1	179.3	1.5	72.9	6.6	6.7	9.6	-177.8	6.2
r5-17	*9*	*-106.9*	*16.5*	*1.9*	*8.0*	*179.5*	*2.6*	*72.8*	*6.8*	*7.4*	*7.9*	*-178.9*	*7.3*
r5-18	9	-132.3	13.7	146.3	17.1	178.9	5.5	53.0	5.5	41.9	9.4	174.8	4.5
r5-18	*3*	*-141.1*	*12.3*	*158.0*	*5.2*	*176.0*	*6.8*	*60.2*	*2.7*	*31.3*	*7.0*	*175.7*	*3.8*
r5-19	8	100.0	13.6	-12.9	10.3	180.0	2.7	-88.1	26.7	134.2	15.6	-179.3	5.3
r5-19	*8*	*91.3*	*13.2*	*-0.6*	*13.9*	*177.7*	*3.6*	*-81.9*	*30.3*	*131.1*	*11.3*	*-178.4*	*4.3*
r5-20	9	-155.4	31.9	151.1	20.8	-176.9	4.7	-60.6	7.9	-30.8	11.4	179.1	4.2
r5-20	*5*	*-157.9*	*31.0*	*155.5*	*17.8*	*179.1*	*3.0*	*-58.9*	*7.8*	*-27.3*	*10.6*	*-178.0*	*3.1*
r5-21	9	84.4	9.7	4.4	11.5	179.4	6.5	-128.5	13.4	149.1	10.6	176.0	2.7
r5-21	*8*	*78.2*	*6.0*	*7.5*	*11.5*	*177.5*	*5.2*	*-125.7*	*17.5*	*149.6*	*9.5*	*176.8*	*3.0*
r5-22	6	-94.1	15.2	-60.1	25.1	178.4	3.5	-176.7	25.0	155.7	12.9	172.1	4.5
r5-22	*3*	*-104.9*	*0.8*	*-72.2*	*10.5*	*175.0*	*4.2*	*-167.8*	*3.7*	*161.0*	*12.8*	*170.5*	*5.5*

Table 4.3.13 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 5 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	φ_{i+2}		ψ_{i+2}		ω_{i+2}		φ_{i+3}		ψ_{i+3}		ω_{i+3}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r5-1	1,902	-89.6	11.6	3.8	10.6	178.3	4.5	85.5	13.1	7.6	14.5	177.6	4.4
r5-1	*1,272*	*-89.4*	*11.7*	*4.0*	*10.4*	*178.4*	*4.3*	*85.5*	*13.3*	*7.2*	*14.7*	*177.7*	*4.4*
r5-2	170	-60.6	8.1	-24.5	12.1	-177.3	3.9	-118.0	12.0	5.3	18.6	-175.8	5.6
r5-2	*101*	*-60.4*	*7.9*	*-24.9*	*11.9*	*-177.5*	*3.8*	*-119.2*	*11.0*	*5.9*	*18.3*	*-176.1*	*5.2*
r5-3	138	68.4	15.1	24.7	15.9	176.4	5.1	70.8	13.8	22.9	16.9	178.7	4.2
r5-3	*80*	*67.8*	*13.7*	*23.9*	*14.3*	*177.6*	*4.2*	*69.2*	*12.7*	*24.5*	*15.4*	*178.3*	*4.1*
r5-4	82	-93.9	19.2	-36.5	16.0	-172.5	4.7	-108.3	15.2	-3.0	12.6	-178.3	5.1
r5-4	*64*	*-92.6*	*17.1*	*-37.7*	*14.1*	*-173.7*	*4.2*	*-104.2*	*15.7*	*-6.5*	*13.1*	*-178.0*	*4.9*
r5-5	69	75.1	12.2	3.1	13.2	-177.8	5.3	-127.0	10.2	-47.4	18.8	-178.0	4.6
r5-5	*47*	*74.2*	*12.0*	*3.9*	*14.5*	*-178.0*	*4.0*	*-128.3*	*12.1*	*-46.3*	*24.1*	*-178.5*	*5.1*
r5-6	67	-55.0	6.1	123.9	6.8	178.9	3.6	73.2	11.7	11.0	12.5	-179.4	4.8
r5-6	*50*	*-56.3*	*6.6*	*124.1*	*6.7*	*179.0*	*3.9*	*75.5*	*11.6*	*8.6*	*10.5*	*-178.9*	*4.7*
r5-7	61	63.8	10.6	14.4	16.0	-174.2	4.9	-129.9	14.7	-32.0	24.4	-176.9	5.5
r5-7	*34*	*62.4*	*7.4*	*13.9*	*14.0*	*-173.6*	*4.2*	*-128.6*	*14.1*	*-31.3*	*23.3*	*-176.9*	*5.3*
r5-8	56	-92.3	18.6	-11.0	21.9	-172.3	6.0	-117.4	24.6	-12.0	23.3	-176.8	5.9
r5-8	*32*	*-91.0*	*20.1*	*-13.7*	*22.8*	*-173.7*	*5.6*	*-113.2*	*22.6*	*-10.4*	*23.0*	*-177.3*	*4.8*
r5-9	29	56.5	8.2	32.2	12.2	179.3	2.6	78.9	11.2	7.4	14.0	-179.9	4.9
r5-9	*19*	*55.3*	*5.9*	*33.4*	*9.2*	*178.8*	*4.3*	*77.9*	*13.7*	*9.3*	*14.4*	*179.4*	*4.2*
r5-10	26	-88.2	15.0	4.6	17.6	179.4	4.2	110.7	18.1	150.9	10.8	-178.8	3.1
r5-10	*19*	*-89.7*	*14.2*	*5.2*	*17.4*	*178.3*	*4.7*	*112.4*	*17.3*	*150.1*	*11.8*	*-178.1*	*5.0*
r5-11	17	-70.2	10.1	-10.0	11.1	177.0	3.3	-84.8	12.2	-4.8	15.9	178.0	5.9
r5-11	*13*	*-73.1*	*13.6*	*-9.4*	*12.4*	*177.8*	*4.3*	*-83.2*	*11.7*	*-8.5*	*20.4*	*-179.7*	*4.9*
r5-12	15	-123.1	12.6	20.4	10.3	179.2	4.5	94.3	13.0	4.2	12.2	179.3	6.4
r5-12	*6*	*-128.1*	*16.3*	*21.9*	*13.9*	*178.5*	*2.3*	*91.7*	*7.9*	*2.7*	*9.4*	*-179.9*	*1.9*
r5-13	16	51.2	4.6	-122.8	6.6	-179.3	2.3	-96.2	8.7	13.0	12.5	-177.5	4.5
r5-13	*12*	*50.6*	*3.9*	*-121.6*	*5.3*	*-179.3*	*2.9*	*-98.4*	*11.8*	*14.6*	*13.6*	*-177.3*	*4.0*
r5-14	12	66.3	10.9	19.0	15.2	179.3	4.9	-136.9	14.3	-153.8	10.0	-178.1	3.1
r5-14	*7*	*66.0*	*13.0*	*19.2*	*17.1*	*-180.0*	*2.8*	*-137.6*	*15.9*	*-156.5*	*9.4*	*-179.1*	*1.5*
r5-15	11	-97.9	5.2	14.2	15.0	-173.8	5.8	-84.9	15.8	-38.1	15.6	178.6	5.8
r5-15	*9*	*-95.7*	*9.3*	*15.2*	*15.4*	*-173.8*	*5.0*	*-81.9*	*7.4*	*-41.7*	*11.8*	*-178.1*	*6.3*
r5-16	9	-95.1	13.8	-30.3	21.9	-170.8	7.9	-115.5	19.1	-32.4	17.4	-177.5	4.4
r5-16	*5*	*-100.6*	*8.8*	*-25.4*	*24.4*	*-176.1*	*4.2*	*-115.0*	*20.1*	*-33.6*	*21.0*	*179.6*	*3.6*
r5-17	9	-130.9	12.8	-64.8	23.5	-176.4	5.6	-117.3	15.7	-16.7	22.2	177.9	8.5
r5-17	*9*	*-133.8*	*9.1*	*-66.2*	*21.8*	*-177.4*	*6.5*	*-112.7*	*19.3*	*-14.7*	*22.9*	*176.0*	*6.6*
r5-18	9	81.9	9.0	-52.3	9.9	-171.0	7.1	-129.7	14.1	4.3	17.1	179.8	5.8
r5-18	*3*	*80.0*	*6.6*	*-44.8*	*9.7*	*-174.9*	*2.9*	*-135.6*	*5.0*	*8.7*	*9.6*	*178.9*	*5.1*
r5-19	8	56.7	5.5	27.1	6.3	177.3	6.2	82.6	13.1	8.4	9.3	-179.3	2.1
r5-19	*8*	*53.3*	*5.5*	*31.6*	*9.1*	*178.1*	*2.4*	*78.5*	*12.1*	*9.1*	*12.8*	*-179.8*	*2.3*
r5-20	9	-103.5	10.3	12.8	10.3	-180.0	4.3	-135.2	15.5	-134.2	10.2	-179.1	4.2
r5-20	*5*	*-107.7*	*5.9*	*8.3*	*6.8*	*179.7*	*1.8*	*-134.1*	*8.1*	*-136.4*	*13.0*	*-178.8*	*2.1*
r5-21	9	51.4	5.3	-129.1	4.6	-178.9	3.8	-74.7	14.0	-8.3	12.1	179.9	4.2
r5-21	*8*	*50.1*	*4.5*	*-126.8*	*4.8*	*-179.3*	*4.4*	*-81.0*	*16.0*	*-4.2*	*15.1*	*-178.1*	*5.6*
r5-22	6	55.9	6.8	30.3	10.1	-179.7	3.2	83.1	12.3	-2.3	7.2	177.5	3.2
r5-22	*3*	*48.6*	*4.6*	*37.7*	*0.5*	*179.9*	*2.9*	*82.7*	*7.4*	*-4.9*	*6.0*	*176.3*	*5.1*

Table 4.3.13 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 5 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	φ_{i+4}		ψ_{i+4}		E_{hb} [kcal/mol]	
		∅	±	∅	±	∅	±
r5-1	1,902	-86.1	17.0	145.2	14.4	-4.8	1.4
r5-1	*1,272*	*-86.0*	*17.3*	*145.7*	*14.6*	*-4.9*	*1.3*
r5-2	170	-134.8	13.9	128.6	16.4	-3.7	1.5
r5-2	*101*	*-133.6*	*13.8*	*128.1*	*16.5*	*-3.9*	*1.3*
r5-3	138	-90.0	20.9	146.0	22.6	-4.5	1.6
r5-3	*80*	*-90.5*	*22.5*	*146.0*	*22.3*	*-4.6*	*1.6*
r5-4	82	77.3	13.3	-143.2	23.6	-3.1	1.3
r5-4	*64*	*78.7*	*12.3*	*-145.7*	*22.5*	*-2.8*	*1.2*
r5-5	69	165.9	49.0	171.5	36.3	-4.4	1.8
r5-5	*47*	*155.0*	*50.4*	*-179.5*	*35.1*	*-4.3*	*1.7*
r5-6	67	-129.7	11.7	118.4	18.4	-4.0	1.4
r5-6	*50*	*-128.4*	*11.3*	*116.5*	*19.9*	*-3.9*	*1.4*
r5-7	61	-120.9	24.6	137.2	20.5	-5.0	1.4
r5-7	*34*	*-123.3*	*25.7*	*136.1*	*20.8*	*-5.0*	*1.6*
r5-8	56	-124.1	25.1	143.1	24.3	-3.9	1.6
r5-8	*32*	*-127.7*	*23.9*	*142.6*	*23.4*	*-3.7*	*1.5*
r5-9	29	-96.0	15.9	141.0	14.5	-4.5	1.3
r5-9	*19*	*-98.1*	*17.6*	*143.6*	*12.1*	*-4.6*	*1.4*
r5-10	26	82.8	15.9	-177.5	20.2	-4.2	1.5
r5-10	*19*	*81.0*	*13.9*	*-173.7*	*22.1*	*-3.7*	*1.5*
r5-11	17	91.3	23.2	-158.1	22.3	-4.5	1.7
r5-11	*13*	*93.7*	*26.5*	*-157.9*	*24.0*	*-4.1*	*1.8*
r5-12	15	-121.1	16.6	157.4	20.6	-3.6	1.2
r5-12	*6*	*-121.6*	*16.1*	*163.1*	*20.8*	*-3.4*	*1.6*
r5-13	16	-98.5	10.8	129.7	8.7	-4.3	1.2
r5-13	*12*	*-100.8*	*10.9*	*129.5*	*9.4*	*-4.7*	*1.0*
r5-14	12	-80.5	10.7	158.2	17.6	-4.4	1.8
r5-14	*7*	*-77.4*	*8.5*	*158.4*	*18.1*	*-3.5*	*2.0*
r5-15	11	-144.5	12.5	129.3	17.7	-4.0	1.8
r5-15	*9*	*-142.0*	*13.5*	*128.4*	*19.7*	*-3.9*	*1.8*
r5-16	9	-154.3	12.1	91.2	20.1	-2.2	1.3
r5-16	*5*	*-156.0*	*8.5*	*87.3*	*22.2*	*-1.6*	*1.0*
r5-17	9	-128.7	23.1	123.8	16.9	-3.2	1.1
r5-17	*9*	*-129.5*	*18.4*	*118.1*	*17.0*	*-3.4*	*1.0*
r5-18	9	-96.0	10.2	138.1	15.9	-3.5	1.5
r5-18	*3*	*-93.4*	*1.7*	*123.5*	*0.9*	*-4.5*	*0.4*
r5-19	8	-95.6	22.5	130.2	14.8	-4.2	1.8
r5-19	*8*	*-89.5*	*20.9*	*126.2*	*12.5*	*-3.2*	*2.1*
r5-20	9	-87.1	14.9	139.2	9.8	-2.8	1.4
r5-20	*5*	*-81.5*	*14.7*	*139.6*	*7.0*	*-3.0*	*1.0*
r5-21	9	-102.9	10.9	51.0	40.2	-2.3	1.3
r5-21	*8*	*-99.9*	*13.8*	*71.9*	*41.2*	*-2.1*	*1.2*
r5-22	6	-80.4	4.7	129.5	17.3	-5.3	1.1
r5-22	*3*	*-83.4*	*7.5*	*141.7*	*8.9*	*-5.9*	*0.7*

Reverse 6 Turns

The number of reverse 6 turn types is also higher as compared to the classification with Secbase. Again, the new classification leads to an aggregation of former turn types, e.g., types IIa and IIb, as well as a separation of more distant turn types, e.g., type IX (see Figure 4.3.19). Furthermore, additional types were found due to the extension of the training set size and the larger number of neurons in the maps.

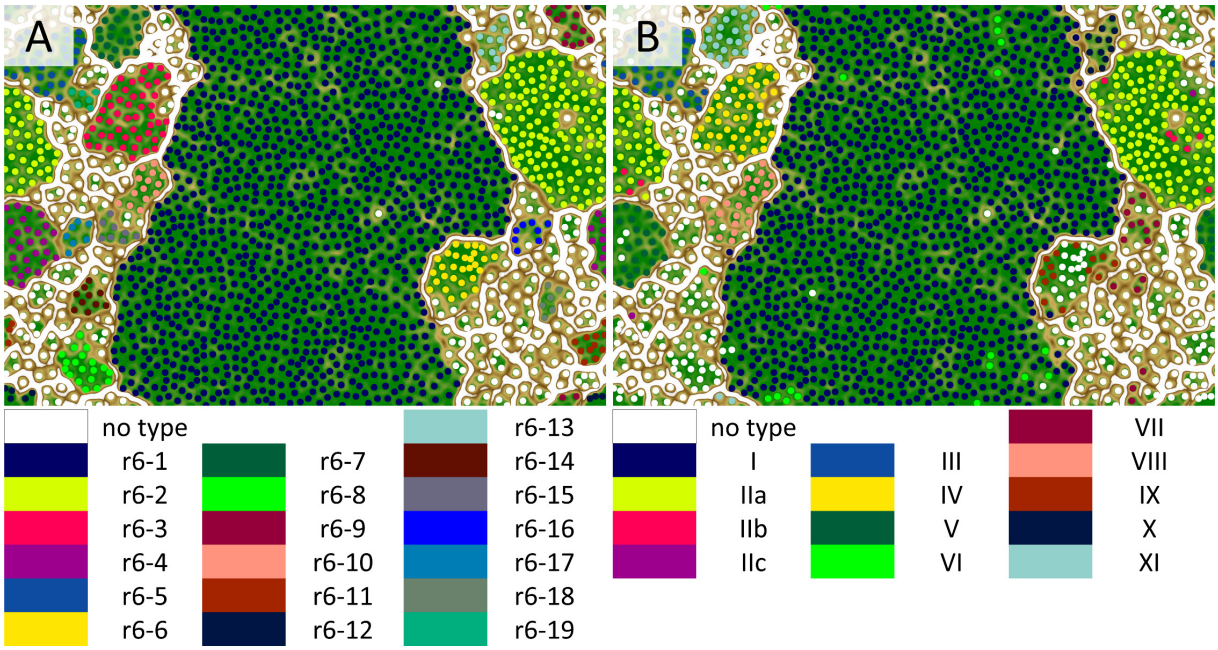


Figure 4.3.19. normal 3 turn clustering. The U-Matrix for reverse 6 turns with the projected best matches of the training set coloured by their turn type assignment in this study (A) and the original Secbase-derived turn classification (B) is shown.

The former types I and VI are assigned to one type called r6-1. This is reasonable when examining the mean dihedral angles of both types in one cluster (Table 4.3.14). As expected, the types IIa and IIb are now part of one unique type r6-2. The turns of type r6-7 which is characterised by a *cis* peptide bond between the residues at the positions *i+1* and *i+2* were formerly characterised as type XI. The turns of types r6-9 and r6-12 with a *cis* peptide bond between residues *i+2* and *i+3* were not assigned by Secbase (no type). Further correspondences for the highest occupied reverse 6 types are: r6-3 to IV, r6-4 to V, r6-5 to III, and r6-6 to IX.

The highest occupied types r6-1 and r6-2 are characterised by an overrepresentation of Gly residues at the positions *i+2* and *i+3*, respectively (see Table A3.19 in Appendix 3 for the over- and underrepresented residues in the most frequent reverse 6 turns). They lead to the strongly bent structure of both turn types which are commonly located between two adjacent β-strands (Table 4.3.12).[431] As already shown previously,[432] these six residue long turn types frequently encompass normal turns of four and five residues leading to the stabilisation of the structures belonging to distinct types. Additionally, normal 6 turns are found inside reverse 6 turns (r6-1 and n6-1, r6-4 and n6-3, see Figure A3.2). The lower number of types as compared to the reverse 5 turns results from the smaller number of hydrogen-bonded turns composed of six residues in the dataset (see also Table 3.3.1).

Table 4.3.14. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 6 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	φ_i		ψ_i		ω_i		φ_{i+1}		ψ_{i+1}		ω_{i+1}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r6-1	1,141	-92.8	22.6	114.0	15.1	-176.3	4.5	-66.0	8.0	-29.3	12.8	-178.9	4.1
r6-1	*762*	*-92.9*	*22.7*	*113.6*	*15.2*	*-176.3*	*4.5*	*-66.0*	*8.0*	*-28.9*	*13.1*	*-179.0*	*3.7*
r6-2	168	-70.7	8.6	161.5	11.9	-177.3	4.2	-59.1	6.5	-30.2	10.5	178.9	3.3
r6-2	*91*	*-70.8*	*9.3*	*163.8*	*12.2*	*-176.9*	*4.0*	*-59.9*	*6.1*	*-28.9*	*10.3*	*178.5*	*4.0*
r6-3	49	-75.7	16.3	129.5	14.9	-177.3	4.2	-55.3	8.0	134.9	8.3	178.0	4.1
r6-3	*30*	*-77.1*	*17.2*	*131.0*	*13.2*	*-177.3*	*3.8*	*-55.9*	*7.7*	*134.6*	*9.6*	*177.8*	*3.8*
r6-4	37	-127.3	13.1	166.6	11.6	176.3	3.4	53.9	6.2	51.3	9.9	178.2	3.0
r6-4	*19*	*-128.6*	*15.6*	*165.9*	*13.6*	*176.4*	*4.0*	*53.8*	*4.8*	*52.0*	*9.3*	*178.3*	*2.1*
r6-5	33	-68.7	7.8	142.6	9.7	-179.4	4.7	-76.3	12.8	-24.4	14.7	-177.0	3.9
r6-5	*19*	*-68.5*	*6.7*	*143.9*	*9.7*	*180.0*	*3.0*	*-78.2*	*12.8*	*-22.7*	*16.0*	*-177.6*	*3.5*
r6-6	29	-87.0	8.2	-177.8	9.2	179.0	4.7	-74.2	7.8	-15.8	15.8	176.9	4.5
r6-6	*21*	*-90.3*	*8.7*	*-175.9*	*10.5*	*179.5*	*3.3*	*-73.2*	*7.6*	*-18.1*	*12.3*	*179.0*	*3.1*
r6-7	24	-83.2	17.0	134.9	13.7	172.3	5.4	-99.8	17.0	135.1	13.2	9.4	12.2
r6-7	*17*	*-80.3*	*16.4*	*133.9*	*15.6*	*174.3*	*4.4*	*-101.1*	*13.4*	*135.0*	*14.6*	*7.0*	*8.2*
r6-8	19	-133.8	11.7	145.4	10.5	173.4	6.3	79.5	7.6	178.4	11.2	-175.3	5.6
r6-8	*11*	*-136.6*	*12.4*	*143.6*	*9.3*	*172.4*	*5.6*	*80.6*	*3.8*	*179.0*	*10.6*	*-173.9*	*4.6*
r6-9	15	-122.7	31.2	144.2	10.5	-179.9	5.0	-92.2	11.3	-36.0	16.9	-178.9	2.6
r6-9	*8*	*-127.4*	*31.2*	*143.1*	*10.3*	*-179.9*	*3.4*	*-91.5*	*12.9*	*-37.3*	*9.4*	*-178.5*	*2.7*
r6-10	14	-76.4	14.6	134.9	14.6	177.3	4.2	-64.6	9.0	145.6	10.5	177.8	5.3
r6-10	*7*	*-81.7*	*21.4*	*128.8*	*29.4*	*178.3*	*5.3*	*-62.8*	*5.6*	*142.6*	*12.3*	*179.7*	*6.9*
r6-11	14	81.1	12.2	1.1	9.3	-178.7	4.2	-125.4	9.8	154.4	11.1	177.0	3.6
r6-11	*12*	*82.7*	*10.1*	*2.0*	*8.9*	*-178.0*	*5.3*	*-129.6*	*9.1*	*155.0*	*10.3*	*175.5*	*3.6*
r6-12	13	-142.4	11.7	151.1	11.4	171.2	3.7	103.4	11.4	175.6	3.4	175.3	4.2
r6-12	*13*	*-143.6*	*12.9*	*151.5*	*11.3*	*173.8*	*3.2*	*99.5*	*10.4*	*176.8*	*5.4*	*177.9*	*3.0*
r6-13	11	-99.1	21.6	130.8	13.1	176.2	4.5	-65.3	6.6	-24.4	14.9	175.8	3.3
r6-13	*8*	*-103.3*	*17.7*	*138.0*	*11.3*	*175.5*	*4.6*	*-66.1*	*7.1*	*-26.0*	*19.9*	*176.6*	*4.0*
r6-14	10	-104.3	16.6	-12.6	18.3	178.5	5.0	-153.4	9.0	173.5	11.9	-179.1	3.8
r6-14	*4*	*-103.1*	*22.7*	*-20.5*	*16.8*	*174.0*	*5.3*	*-154.9*	*8.3*	*176.3*	*18.0*	*-179.7*	*6.3*
r6-15	8	-92.2	19.4	109.2	11.1	-172.5	6.1	-58.7	6.0	125.5	9.4	179.8	3.1
r6-15	*3*	*-104.4*	*5.8*	*98.7*	*8.7*	*-172.9*	*5.7*	*-60.0*	*7.2*	*127.2*	*3.7*	*-177.5*	*3.4*
r6-16	8	-98.9	8.0	-129.9	20.8	-177.4	3.9	-74.3	10.2	-21.9	15.5	177.8	5.2
r6-16	*8*	*-100.6*	*8.7*	*-129.7*	*28.7*	*-177.7*	*4.4*	*-75.3*	*13.4*	*-14.5*	*19.1*	*179.5*	*13.0*
r6-17	7	-88.5	10.8	165.6	5.9	178.0	1.4	-54.7	6.5	139.6	6.2	178.6	4.7
r6-17	*4*	*-85.8*	*7.3*	*162.8*	*5.8*	*178.0*	*2.3*	*-54.3*	*1.6*	*134.5*	*9.3*	*179.9*	*3.4*
r6-18	6	68.7	8.1	23.1	19.3	179.3	3.4	-88.3	17.0	-12.7	15.8	177.3	3.0
r6-18	*4*	*64.1*	*7.1*	*30.9*	*11.2*	*176.3*	*4.2*	*-96.7*	*12.9*	*-1.4*	*3.9*	*179.3*	*2.1*
r6-19	6	-65.2	5.3	146.1	9.9	177.8	2.3	-74.6	20.3	-26.1	12.2	-173.5	4.8
r6-19	*4*	*-64.5*	*7.5*	*141.0*	*4.5*	*-179.9*	*1.9*	*-75.1*	*15.6*	*-15.5*	*10.4*	*-176.3*	*2.3*

Table 4.3.14 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 6 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	φ_{i+2}		ψ_{i+2}		ω_{i+2}		φ_{i+3}		ψ_{i+3}		ω_{i+3}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r6-1	1,141	-74.6	13.2	-40.5	12.3	-177.9	4.2	-102.0	16.3	-9.3	11.6	-177.9	4.8
r6-1	*762*	*-74.6*	*13.1*	*-40.3*	*12.3*	*-178.1*	*4.0*	*-102.3*	*16.3*	*-9.2*	*11.8*	*-178.1*	*4.9*
r6-2	168	-84.7	11.1	-2.3	10.1	178.7	4.1	70.3	9.7	14.1	14.4	-177.2	4.9
r6-2	*91*	*-86.1*	*12.5*	*-2.6*	*10.1*	*178.6*	*4.2*	*70.0*	*9.1*	*15.4*	*12.7*	*-177.5*	*4.9*
r6-3	49	70.4	12.1	16.2	14.6	179.2	4.5	-139.9	13.2	168.8	10.8	175.9	5.7
r6-3	*30*	*69.9*	*12.4*	*18.2*	*16.9*	*179.9*	*4.2*	*-140.7*	*15.1*	*170.8*	*10.4*	*176.9*	*6.0*
r6-4	37	59.0	7.5	33.4	11.2	179.0	3.6	86.7	17.3	7.5	13.2	-179.8	5.0
r6-4	*19*	*60.0*	*6.9*	*32.9*	*8.8*	*179.1*	*3.1*	*85.0*	*16.1*	*7.8*	*12.2*	*-179.2*	*3.9*
r6-5	33	-113.1	18.7	-25.0	18.0	-179.6	5.3	-120.0	27.0	155.9	7.8	179.9	3.6
r6-5	*19*	*-112.0*	*15.0*	*-25.6*	*18.7*	*-179.0*	*5.3*	*-121.6*	*25.7*	*154.0*	*7.5*	*178.3*	*3.0*
r6-6	29	-98.8	13.6	128.8	12.4	-178.9	4.0	-55.5	4.0	136.1	5.2	178.2	2.6
r6-6	*21*	*-98.8*	*11.7*	*129.6*	*13.2*	*180.0*	*4.1*	*-54.0*	*3.8*	*134.6*	*6.5*	*178.4*	*2.1*
r6-7	24	-96.5	8.3	-4.7	8.1	-172.8	4.6	-77.7	11.6	-21.8	10.5	-179.8	5.6
r6-7	*17*	*-96.2*	*8.5*	*-2.4*	*7.9*	*-171.9*	*4.6*	*-76.3*	*7.2*	*-22.6*	*8.8*	*-178.4*	*5.8*
r6-8	19	-64.4	7.8	-27.3	9.7	179.2	3.4	-95.6	11.6	4.5	10.0	177.7	4.8
r6-8	*11*	*-64.5*	*5.7*	*-27.7*	*10.5*	*179.0*	*3.6*	*-96.1*	*13.8*	*3.5*	*9.1*	*177.3*	*4.0*
r6-9	15	-135.2	7.3	103.9	10.2	-4.5	4.9	-74.1	5.8	164.4	8.5	-178.1	2.8
r6-9	*8*	*-133.7*	*8.7*	*104.8*	*9.7*	*-4.3*	*4.3*	*-73.2*	*7.5*	*163.7*	*8.9*	*-178.2*	*1.6*
r6-10	14	79.3	9.3	-59.0	9.0	179.9	5.1	-80.2	9.6	-24.0	7.1	179.2	6.4
r6-10	*7*	*81.6*	*11.8*	*-58.0*	*13.1*	*178.3*	*6.5*	*-80.5*	*10.9*	*-22.2*	*9.2*	*176.5*	*8.5*
r6-11	14	52.3	4.8	-125.4	7.0	-178.5	2.7	-94.5	10.2	3.8	8.8	176.6	4.3
r6-11	*12*	*53.4*	*5.8*	*-126.1*	*8.9*	*-177.8*	*3.4*	*-91.4*	*12.3*	*0.7*	*9.1*	*178.0*	*6.7*
r6-12	13	-129.6	5.6	107.9	7.6	-2.7	4.2	-72.9	5.2	160.9	5.8	-175.4	2.8
r6-12	*13*	*-130.5*	*8.5*	*105.8*	*9.4*	*-0.8*	*3.9*	*-74.0*	*7.4*	*160.2*	*7.0*	*-175.5*	*3.4*
r6-13	11	-78.6	13.6	104.4	13.2	-178.8	1.7	85.5	12.7	-4.4	17.0	-178.5	5.5
r6-13	*8*	*-81.1*	*20.2*	*113.3*	*20.2*	*-179.6*	*1.7*	*78.5*	*7.0*	*4.4*	*10.9*	*-179.7*	*9.4*
r6-14	10	-74.1	8.9	-5.1	9.1	176.4	2.5	-89.6	12.1	-0.9	13.0	176.9	4.0
r6-14	*4*	*-68.3*	*5.5*	*-19.3*	*2.0*	*176.2*	*2.8*	*-74.8*	*3.3*	*-11.9*	*9.0*	*175.1*	*4.1*
r6-15	8	76.5	15.8	3.8	12.3	-179.9	7.0	-110.5	13.3	-8.6	9.4	177.9	5.8
r6-15	*3*	*82.5*	*10.5*	*-6.1*	*3.6*	*-177.4*	*3.0*	*-102.0*	*5.8*	*-5.8*	*4.5*	*176.2*	*4.9*
r6-16	8	-142.8	20.1	166.2	11.7	-177.3	5.1	-65.5	8.8	-12.3	12.5	179.4	5.1
r6-16	*8*	*-148.3*	*17.1*	*160.6*	*15.7*	*-176.5*	*5.8*	*-67.5*	*10.2*	*-11.1*	*13.4*	*178.4*	*5.9*
r6-17	7	95.9	17.8	-8.2	12.2	179.2	3.3	61.8	6.4	16.1	17.5	-177.7	5.6
r6-17	*4*	*84.6*	*21.7*	*-0.4*	*21.6*	*177.3*	*2.2*	*63.8*	*4.9*	*29.2*	*5.9*	*-179.4*	*7.6*
r6-18	6	-161.8	18.7	172.1	4.3	-179.5	3.1	-61.9	8.5	-33.1	13.4	179.3	3.7
r6-18	*4*	*-171.6*	*13.8*	*172.6*	*2.0*	*-178.3*	*2.3*	*-60.6*	*6.2*	*-39.8*	*6.8*	*-179.9*	*1.4*
r6-19	6	-108.4	19.6	-8.4	10.8	179.2	4.5	-143.8	6.3	178.8	7.2	-179.6	6.1
r6-19	*4*	*-116.7*	*9.1*	*-1.5*	*2.2*	*178.4*	*6.0*	*-149.5*	*5.8*	*-179.2*	*8.4*	*-177.8*	*5.7*

Table 4.3.14 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and hydrogen bond energies of the classified turn types for the training and test set of reverse 6 turn structures. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	φ_{i+4}		ψ_{i+4}		ω_{i+4}		φ_{i+5}		ψ_{i+5}		E_{hb} [kcal/mol]	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r6-1	1,141	65.6	12.4	27.7	20.5	179.7	4.5	-105.2	30.4	146.0	15.1	-5.5	1.3
r6-1	*762*	*66.0*	*12.6*	*26.7*	*20.7*	*179.8*	*4.3*	*-105.5*	*30.3*	*146.4*	*15.1*	*-5.6*	*1.3*
r6-2	168	-96.2	18.3	-21.6	14.2	-178.9	5.0	-131.3	22.3	146.6	16.6	-5.0	1.4
r6-2	*91*	*-98.9*	*20.1*	*-21.6*	*15.3*	*-179.3*	*5.3*	*-130.3*	*23.1*	*145.6*	*16.4*	*-5.0*	*1.3*
r6-3	49	-97.4	9.9	15.8	21.7	-178.8	4.1	-92.7	22.3	136.9	16.4	-5.1	1.1
r6-3	*30*	*-97.8*	*11.3*	*11.7*	*22.4*	*-178.3*	*3.4*	*-91.1*	*24.4*	*136.8*	*16.9*	*-4.8*	*1.2*
r6-4	37	-82.4	14.6	-17.8	10.6	178.5	5.1	-135.3	12.4	140.1	17.6	-4.8	1.3
r6-4	*19*	*-85.6*	*12.9*	*-14.8*	*8.9*	*177.7*	*4.7*	*-135.8*	*15.1*	*137.5*	*18.6*	*-5.1*	*1.4*
r6-5	33	-81.7	5.0	70.6	13.5	-177.8	3.6	-139.9	17.0	148.1	16.6	-4.9	1.6
r6-5	*19*	*-79.9*	*5.7*	*73.4*	*13.2*	*-178.7*	*5.2*	*-142.2*	*13.9*	*147.2*	*14.9*	*-5.3*	*1.5*
r6-6	29	80.6	8.9	1.8	10.3	178.7	3.7	-113.0	11.2	133.9	14.5	-3.9	1.5
r6-6	*21*	*78.5*	*10.8*	*6.3*	*12.3*	*177.9*	*4.0*	*-114.6*	*11.3*	*132.9*	*18.0*	*-4.1*	*1.6*
r6-7	24	77.3	14.6	17.0	20.2	177.9	5.4	-78.7	12.6	146.7	15.6	-4.9	1.4
r6-7	*17*	*78.6*	*13.3*	*14.5*	*20.7*	*179.3*	*4.4*	*-75.6*	*12.2*	*142.6*	*10.7*	*-4.6*	*1.4*
r6-8	19	54.2	7.6	53.2	9.9	177.3	3.5	-78.7	15.3	135.9	10.6	-2.8	0.9
r6-8	*11*	*56.4*	*7.9*	*51.1*	*11.8*	*178.6*	*3.5*	*-74.8*	*9.6*	*137.9*	*6.8*	*-2.7*	*1.0*
r6-9	15	-72.0	7.5	113.6	11.3	179.4	5.2	-144.0	18.8	146.4	17.0	-3.6	2.1
r6-9	*8*	*-71.7*	*6.4*	*113.2*	*11.8*	*179.8*	*4.6*	*-138.7*	*9.3*	*147.5*	*17.4*	*-3.0*	*2.1*
r6-10	14	67.3	10.4	22.8	13.2	179.2	2.4	-76.2	14.3	133.6	9.0	-5.0	1.1
r6-10	*7*	*69.7*	*12.1*	*22.1*	*14.7*	*179.4*	*2.6*	*-80.6*	*14.2*	*133.7*	*6.5*	*-4.8*	*1.2*
r6-11	14	-81.4	7.9	155.6	11.0	178.8	2.3	-68.5	13.8	156.3	18.6	-3.5	1.4
r6-11	*12*	*-83.3*	*11.4*	*154.8*	*11.2*	*176.3*	*2.9*	*-68.6*	*13.2*	*159.9*	*19.3*	*-3.8*	*1.7*
r6-12	13	-67.2	5.2	135.6	4.7	174.4	5.6	-123.7	12.0	142.9	13.2	-3.8	0.8
r6-12	*13*	*-65.2*	*5.7*	*138.0*	*4.3*	*173.8*	*4.8*	*-124.6*	*10.3*	*144.7*	*13.3*	*-3.5*	*0.7*
r6-13	11	55.4	14.4	42.4	21.2	-176.6	5.9	-94.1	24.5	137.1	12.8	-5.5	0.8
r6-13	*8*	*51.6*	*8.8*	*46.8*	*15.8*	*-174.6*	*7.6*	*-96.4*	*15.1*	*134.6*	*12.0*	*-5.4*	*0.8*
r6-14	10	83.7	18.4	14.3	14.3	176.6	5.9	-72.1	9.1	141.3	11.0	-5.6	1.5
r6-14	*4*	*89.0*	*16.8*	*20.1*	*6.9*	*171.0*	*7.1*	*-70.2*	*14.3*	*138.6*	*12.3*	*-5.6*	*0.8*
r6-15	8	64.0	11.0	26.9	18.9	-174.6	3.3	-128.8	24.3	145.0	13.1	-5.3	1.1
r6-15	*3*	*64.3*	*6.0*	*27.8*	*13.0*	*-174.8*	*3.7*	*-146.3*	*13.8*	*154.7*	*7.2*	*-5.4*	*1.4*
r6-16	8	-96.4	9.9	1.4	8.9	-178.3	4.0	-119.7	14.2	118.9	20.7	-3.6	1.9
r6-16	*8*	*-94.7*	*8.5*	*-4.1*	*10.2*	*-177.0*	*5.5*	*-121.7*	*12.1*	*129.3*	*8.8*	*-2.5*	*1.2*
r6-17	7	-94.3	26.0	-28.2	12.2	-179.4	3.8	-128.9	18.7	153.6	15.3	-4.0	1.5
r6-17	*4*	*-91.0*	*15.2*	*-15.7*	*15.1*	*178.9*	*2.7*	*-153.9*	*5.5*	*161.1*	*6.8*	*-4.4*	*2.0*
r6-18	6	-71.7	9.7	-16.2	9.0	-179.8	3.6	-119.0	9.8	23.6	22.4	-2.3	0.9
r6-18	*4*	*-66.2*	*1.3*	*-15.6*	*11.1*	*-178.1*	*3.5*	*-116.4*	*12.0*	*11.6*	*17.7*	*-2.5*	*0.9*
r6-19	6	-85.1	9.2	2.8	9.6	-176.6	1.9	-72.2	7.4	147.4	13.0	-5.6	0.8
r6-19	*4*	*-88.1*	*12.5*	*3.1*	*8.9*	*-176.8*	*1.4*	*-76.8*	*5.0*	*146.9*	*10.4*	*-5.8*	*0.4*

4.3.3 Assignment of Helices and Sheets with SCOT

The assignment of helices based on turns is highly different from the SHAFT-derived assignments and does not incorporate special helix-capping motifs. In contrast, the aim was to detect geometrically regular and stable helices and strands which is essential for the accurate modelling of secondary structure arrangements in proteins. A new graph-based method was developed for the classification of sheets. Even the assignment of β-strands uses turn information to split them appropriately. The parsing of the PDB files and implementation of SCOT was realised by Dipl.-Inf. Tobias Brinkjost and will be an integral part of his PhD thesis. He also reimplemented the SHAFT procedure which uses the same pre-processing steps. However, the SHAFT classifications of α-, 3_{10}-, and π-helices are assigned using the Secbase-assigned turn types.

The final classification of helices and sheets by SCOT incorporates the hydrogen bonding patterns, the dihedral angles (via turn types introduced in Section 4.3.2), and also the backbone geometry by analysing Cα-Cα distances within the assigned helices and strands. The optimisation of the final SSE classification was realised by visual inspection and a systematic analysis of the four-residue segment Cα-Cα distances within the SSE classes. The general approaches for the new method will be summarised below and the results are compared to the reimplemented SHAFT methodology to discuss the most significant differences.

α-, 3_{10}-, and π-Helix Assignment

The assignment of new turn classes allowed for novel approaches with respect to the helix extension and capping. Due to the small number of capping motifs which can be captured by analysing the main-chain hydrogen bonds, the search for N- and C-terminal capping motifs was completely omitted in SCOT.[433] After the definition of core helix residues, i.e. the residues which are involved in at least two n5-1 or n4-1 turns for α- and 3_{10}-helices, respectively, the N- and C-terminal positions of α- and 3_{10}-core helices were checked for the presence of open turns. The helices were N-terminally and C-terminally extended as long as a residue was spanned by at least three o4-2, o5-1, or o6-4 turns.

Although typical open turns at specific positions of the N- and C-termini could be identified, their deviations from the optimum helical dihedral angles at the first or last positions are high. This would inevitably lead to backbone distortions at the helix termini. Therefore, we abandoned their use in the helix assignment procedure. The omission of the helix-capping procedures applied in SHAFT led to helix termini with regular φ and ψ angles (see Table 4.3.16). Additionally, the number of already known capping motifs is very large[434] and involves various side-chain interactions[433,435] in addition to the backbone hydrogen bond-stabilised capping motifs.[436,437] A reliable cap definition is only possible by also considering interactions of the side-chain atoms. Therefore, we argue that the extension by open turns is sufficient to assign appropriate Ncap or Ccap residues and reduce the risk of the inclusion of residues with non-helical geometries at the N- and C-termini.

Finally, overlapping α- and 3_{10}-helices are merged according to predefined rules and the class assignment depends on the number of overlaps of the corresponding turn types (see below).

In our SCOT approach, π-helix residues are defined if they are spanned by at least two n6-2 turns. π-helical backbone segments can also be part of α- or 3_{10}-helices. In contrast to the latter ones, π-helices are extended based on o4-6 turn overlaps only, and they are not merged with other right-handed helices as their characteristic *i* to *i+5* hydrogen bonding motif can often be found in combination with *i* to *i+4* and *i* to *i+3* backbone hydrogen bonds. A merging would lead to an insufficient detection of π-helical conformations in protein structures hampering their analysis. Instead of using a hierarchical class assignment, preferring α-helices over 3_{10}- and π-helices, we calculated Purity values based on the number of n4-1, n5-1, and n6-2 turns spanning the residues of the helix. For helices with equal purity values for the three classes, we assign type 0 (mixed helix).

Helical Irregularities

SCOT incorporates two approaches to avoid long highly bent or kinked helices. For α-, 3_{10}-, and π-helices, the four-residue segment Cα-Cα distances were analysed to find a maximum distance that guarantees a regular helical structure. For four-residue segment Cα-Cα distances that are above 6.25, 6.55, and 7 Å for α-, 3_{10}-, and π-helices, respectively, we incorporate a splitting procedure that divides the kinked helix into two parts. This ensures low bending angles within helical regions. In a second step, we look for disruptions of the regular hydrogen bonding pattern that is usually observed for the helix classes and assign helix kink residues whenever the number of hydrogen bonds is significantly lower than expected. This procedure identifies regions in helices which are less stabilised. These adjustments are especially helpful for the modelling of SSEs with the help of vector-based representations.

Further Helix Types

The normal turn types n4-3 and n5-9 are used to assign left-handed α- and 3_{10}-helices, respectively. An extension of left-handed helices is possible based on the occurrence of o4-15 or o5-46 turns, but was not realised in the final SCOT methodology. Turns with dihedral angles that are typical for left-handed π-helices could not be found. Like their right-handed counterparts, they are comparatively rare. An increase in the size of the dataset used for the assignment of turn types might lead to the detection of a normal 6 turn type which constitutes left-handed π-helices. Helix purity values based on the number of the helix-spanning n4-3 and n5-9 turns are also calculated for the left-handed helices. Due to the generally very short helices, a splitting procedure was not introduced for this helix category.

PPII helices are assigned based on overlapping o4-9 turns with Cα-Cα distances between 4 and 10 Å. Pro residues are overrepresented in this turn type (see Section 4.3.2). A minimum turn overlap of two over a length of four residues is necessary for the PPII helix assignment with SCOT. Similar to the regular right-handed helices, PPII helices are split if the four-residue segment Cα-Cα distance is below 7.45 Å. Due to its extended conformation, this helix type is referred to as ribbon. However, while 34.5% of the o4-9 turns are found outside the regular SSEs, 15.5% of the o4-9 turns are located inside β-strands leading to the assignment of overlapping PPII helices and strands. Therefore, the assignment of the PPII conformation was restricted to sequence segments which are not assigned as strands.

SCOT supports two further SSE types which can also be classified as ribbon helices. These right- and left-handed 2.2_7-helices (overlapping turns of types n3-2 and n3-1) could rarely be identified with the new hydrogen bond criterion and the corresponding turn detection. This is in line with the small amount of published examples for these SSEs.[438–440] An analysis of the complete PDB led to the classification of two right-handed and ten left-handed 2.2_7-helices. The latter are assigned based on the turn type which corresponds to the one for the assignment of γ-helices with SHAFT. In contrast to SHAFT, the SCOT-assigned left-handed 2.2_7-helices do not overlap with β-strands (see Figure A3.3 in Appendix 3 for two examples) underlining the benefits of the new hydrogen bond criterion.

β-Sheets

Details regarding the β-strand assignment will be an integral part of the doctoral thesis of Dipl-Inf. Tobias Brinkjost. In short, we search for the typical parallel and antiparallel hydrogen bonding patterns of β-sheets. Consequently, the detection of isolated strands as realised by geometry-based SSAMs is not possible. The assignment of bent strands with intra main-chain hydrogen bonds is avoided by the blocking of residue positions based on the presence of reverse 2, 3, and 4 turns and normal 3 and 4 turns. In contrast to DSSP, we additionally apply four-residue segment Cα-Cα distances to identify kinked β-strands (for four-residue segment Cα-Cα distances below 8.5 Å) which can be used to split strands into linear protein backbone segments. However, this splitting procedure was not included as default setting of the method as an additional focus of SCOT is on the consistent assignment of stable SSEs.

Comparison of SCOT and SHAFT Assignments

Besides the fact that turns are the basis for SCOT- and SHAFT-based classifications, both methods differ significantly leading to their unique features discussed below. The major differences are: the hydrogen bond assignment method, the type of data used for the ESOM-based clustering of turn types, the assignment of helices, the omission of helix-capping rules, the additional assignment of β-sheets, the use of geometric criteria to avoid helical irregularities, and the assignment of helix and strand kinks with SCOT.

The SCOT helix assignment is more restrictive with respect to the hydrogen bond definition and the definition of helix termini. This leads to geometrically more consistent helix assignments by SCOT and the omission of nearly all previously identified γ-helices within β-strands. In contrast, SHAFT identifies the usually rare γ-helices in 1,926 of all 3,597 structures of our evaluation dataset. Table 4.3.15 summarises the characteristics of the assigned helices by both methods for our representative dataset of high-quality X-ray structures. It shows that SHAFT-assigned helices are approximately two (3_{10}- and π-helices) or three (α-helices) residues longer. This is the result of the helix merging and capping procedure in the SHAFT method. These approaches were avoided for our purposes in favour of more regular and consistent helix assignments. This difference also explains the large standard deviations for the dihedral angles of residues which are assigned to right-handed helices by SHAFT.

The only exception are the SHAFT-based π-helix assignments which are geometrically more regular than those of SCOT. However, this can be partially attributed to the small number of identified π-helices by SHAFT (47 vs. 572 SCOT-assigned π-helices). Additionally, SCOT assigns π-helices that overlap with other right-handed helix types leading to high deviations from the ideal helical geometry.

The higher number of π-helices identified by SCOT results from the finding that this helix class should be analysed independent of other right-handed helices leading to π-helices overlapping with α- and 3_{10}-helices. Consequently, SCOT-assigned π-helices show very low mean Purities as compared to other helix classes. This relates to the finding that the π-helix is a cryptic structure often referred to as bulge in regular helical structures.[97]

A further step to avoid geometric irregularities in helices was the use of four-residue segment Cα-Cα distances for α-, 3_{10}-, π-, and PPII helices. Kinked helices are characterised by distinct minima of these distances at kink positions which were used to split helices of all four classes. This splitting step leads to lower average bending angles of SCOT-assigned helices as compared to the SHAFT-assigned helices and geometrically more regular helical geometries as reflected by the low standard deviations of the helical twist, rise, and radius.

Table 4.3.15. The geometric parameters of α- (H1), 3_{10}- (H5), π- (H3), and mixed (H0) helices assigned by SCOT and SHAFT. The angular parameters twist, vtor (virtual torsion angle), and the bending angle (BDA) as well as the backbone dihedral angles are given in degrees. The radius (r) and the rise are given in Å. Mean values of all parameters and the respective standard deviations in parentheses are given. The given B-factor was scaled to ensure comparability between multiple protein structures. Purity values for SHAFT-assigned helices were not calculated.

		twist	**rise**	**r**	**vtor**	**BDA**	***φ***	***ψ***	***ω***	**B-factor**	**Length**	**Purity**
H1	SCOT	98.9	1.5	2.3	49.7	5.8	-66.0	-38.0	179.5	-0.1	11.7	0.9
		(3.7)	(0.1)	(0.1)	(5.9)	(4.0)	(11.9)	(13.6)	(4.0)	(0.8)	(5.3)	(0.1)
	SHAFT	98.9	1.5	2.3	50.2	8.3	-67.3	-35.1	179.7	-0.1	14.5	-
		(5.8)	(0.2)	(0.2)	(10.1)	(9.4)	(26.6)	(36.8)	(7.7)	(0.9)	(5.7)	-
H0	SCOT	101.0	1.6	2.2	54.4	13.9	-71.1	-29.5	-179.3	-0.1	5.9	0.5
		(6.7)	(0.2)	(0.2)	(11.6)	(9.5)	(18.9)	(18.9)	(4.7)	(0.9)	(1.8)	(0.0)
H5	SCOT	105.5	1.7	2.1	62.1	13.6	-71.9	-20.4	-179.5	0.1	3.7	0.9
		(7.4)	(0.2)	(0.2)	(12.6)	(7.8)	(19.2)	(17.9)	(4.5)	(0.9)	(1.2)	(0.1)
	SHAFT	104.9	1.7	2.1	62.1	31.4	-77.5	-2.5	-179.4	0.2	5.7	-
		(9.0)	(0.4)	(0.2)	(17.0)	(27.6)	(30.8)	(59.1)	(9.2)	(1.0)	(1.3)	-
H3	SCOT	82.7	1.2	2.7	31.3	14.2	-77.9	-43.2	-176.9	-0.3	6.5	0.6
		(6.7)	(0.4)	(0.3)	(11.8)	(7.5)	(21.1)	(14.6)	(4.8)	(0.7)	(1.7)	(0.1)
	SHAFT	84.5	1.2	2.7	33.3	15.2	-76.8	-41.4	-177.6	-0.3	8.4	-
		(8.5)	(0.4)	(0.3)	(14.1)	(8.3)	(25.0)	(25.0)	(4.8)	(0.7)	(2.8)	-

The difference in the detection of helix termini leads to a shift concerning the over- and underrepresented residues at the helix cap positions, as well as atypical dihedral angles at the helix termini for the SHAFT-assigned helices (Table 4.3.16). The results for π-helices are not presented as the number of SHAFT-assigned π-helices is too low to derive significant statistics. The SCOT-assigned helix termini are geometrically more consistent as observed by the investigation of the corresponding dihedral angles at the Ncap and Ccap positions.

These dissimilarities can be attributed to the differences in the basic helix assignment approaches. While SHAFT was designed to analyse the meaning of SSEs for protein function and their impact on drug design,[93] SCOT was developed as assignment method that enables a geometrically uniform representation of regular and stable backbone conformations for subsequent SSE-based structure comparisons.

Table 4.3.16. The helix Ncap and Ccap assignments by SCOT and SHAFT. The over- and under-represented residues and the mean dihedral angles with standard deviations (stdev) for the SCOT- and SHAFT-assigned α- and 3_{10}-helices at different helix-related positions.

	SCOT			**SHAFT**		
	over-represented (d > 4.4)	***under-represented (d < -4.4)***	***φ±stdev ψ±stdev***	***over-represented (d > 4.4)***	***under-represented (d < -4.4)***	***φ±stdev ψ±stdev***
$Ncap_{-1}$	DGNPST	AEFIKLMQ RVWYX	-	GILMP	AEHKQRX	-
Ncap	AELPW	CDGHNST	-59.1°±8.7° -40.1°±9.7°	DGNPST	AEFIKLMQ RVWYX	-93.5°±45.2° 131.0°±53.2°
$Ncap_{+1}$	ADEPQ	CFGHILMN TVYX	-	EPW	CDGHINST VX	-
α-helix	ARQEILKMX	NDCGHPST V	-66.0°±11.9° -38.0°±13.6°	ARQELKM	NCGHIPST YV	-67.3°±26.6° -35.1°±36.8°
$Ccap_{-1}$	AEKLQR	CFGHIPTVW YX	-	AFHKLNQRTY	DGIPVX	-
Ccap	AFHKLNQRTY	DGIPV	-94.1°±18.1° -6.8°±23.5°	GN	AEILMPRT VWY	-50.6°±75.8° 66.3°±60.8°
$Ccap_{+1}$	GN	ADEFILMP RTVWYX	-	DKP	AEFHLMST VWYX	-
$Ncap_{-1}$	DHNPS	AEFILMQVX	-	FIV	AEHQX	-
Ncap	ALPW	DGHNQTV	-57.7°±10.7° -34.2°±11.9°	DHNPS	AEFILMQV	-92.9°±44.7° 122.2°±51.7°
$Ncap_{+1}$	ADEPSW	FGILMTVYX	-	PW	DGHNQTVX	-
3_{10}-helix	DQELKFPSWY	NGITV	-71.9°±19.2° -20.4°±17.9°	DQELKFPSWY	NGITV	-77.5°±30.8° -2.5°±59.1°
$Ccap_{-1}$	DEKQS	FGILTVX	-	DFKLNQY	AGIPVX	-
Ccap	DFKLNQY	AGIPSV	-96.8°±15.9° -2.8°±18.0°	CFILV	ADEKPT	-87.9°±39.2° 126.8°±52.2°
$Ccap_{+1}$	CFGILV	DEKPSTX	-	DGP	AEILVYX	-

Secondary Structure Elements Not Classified with SHAFT

Another difference to SHAFT is the additional assignment of left-handed helices (α-, 3_{10}-, and PPII helices) and β-strands with SCOT. Consequently, these additional features of SCOT had to be validated by comparing it to SSAMs other than SHAFT. Tables A3.20-22 in Appendix 3 present details concerning the following discussions.

While our definition of β-strands shows a high consensus to that of the widely applied method DSSP (90% consensus), the reliability of the assignment of left-handed helices is more difficult to assess. For PPII helices, all methods which support the assignment of the PPII conformation in protein structures have a considerably low consensus.[113] The same holds true for SCOT as compared to ASSP, DISICL, and SEGNO PPII definitions. Nonetheless, the geometric characteristics of SCOT-assigned PPII helices underline the validity of the new assignment approach. Table A3.20 presents the consensus of all methods. The geometric characteristics of the SCOT-assigned PPII helices are listed in Table A3.21.

In contrast, we could show a high agreement in the assignment of left-handed helices with the geometry-based methods ASSP and DISICL (Table A3.22). However, SCOT assigns the left-handed helix types (α, 3_{10}) more accurately as the hydrogen bonding pattern underlying the classes is used for differentiation purposes. This was shown by an analysis of the dataset of left-handed helices by Novotny and Kleywegt.[90] Another benefit of using turn classes for the assignment of left-handed helices is the independence of the virtual torsion angle, i.e., the dihedral angle which is defined by the four succeeding Cα atoms of a four-residue segment, for the handedness definition. The use of this angle for the helix handedness definition with DSSP leads to multiple one residue long left-handed helices at the Ccap positions of right-handed helices. This can be attributed to the fact that the calculation of the virtual torsion angle also considers the two following residues outside the regular helical structure at the C-terminus.

The main differences between SCOT and three other SSAMs are summarised for one example in Figure 4.3.20. The purely geometry-based method ASSP does not classify a 3_{10}-helix (yellow circle) detected by the other SSAMs due to its high deviations from the optimum helical geometry.[98] A long α-helix as classified by SHAFT is split into several geometrically more uniform helical segments by ASSP, DSSP, and SCOT (red circle). The SHAFT-assigned termini of this helix are also more geometrically irregular than those of the other methods. ASSP and DSSP assign a π-helix at the C-terminus of an α-helix which is not assigned by SCOT and SHAFT. Instead, SHAFT merges both helices, whereas SCOT neglects the single n6-2 turn constituting the π-helix (green circle). The occurrence of n6-2 turns at the termini of other right-handed helix types hints at the presence of further capping motifs (apart from the Schellman motif) at the N- and C-terminus which might be worth further investigations. ASSP, DSSP, and SCOT additionally assign a left-handed helix. DSSP assigns this helix as left-handed α-helix, which is arguable as we calculated a 3_{10}-helical purity of 0.636 (based on the overlaps of the underlying n4-3 turns) and an α-helical purity of 0.364 (based on the overlaps of the n5-9 turns) for this backbone stretch. Finally, SCOT assigns a PPII helix which was not detected by ASSP, while a PPII helix assigned by ASSP is situated in a region which was classified as β-strand by DSSP and SCOT. All these observations are in line with the results observed for the complete dataset and will be discussed in more detail in the doctoral thesis of Dipl.-Inf. Tobias Brinkjost. They hint at the unique improvements in SSE assignments by SCOT.

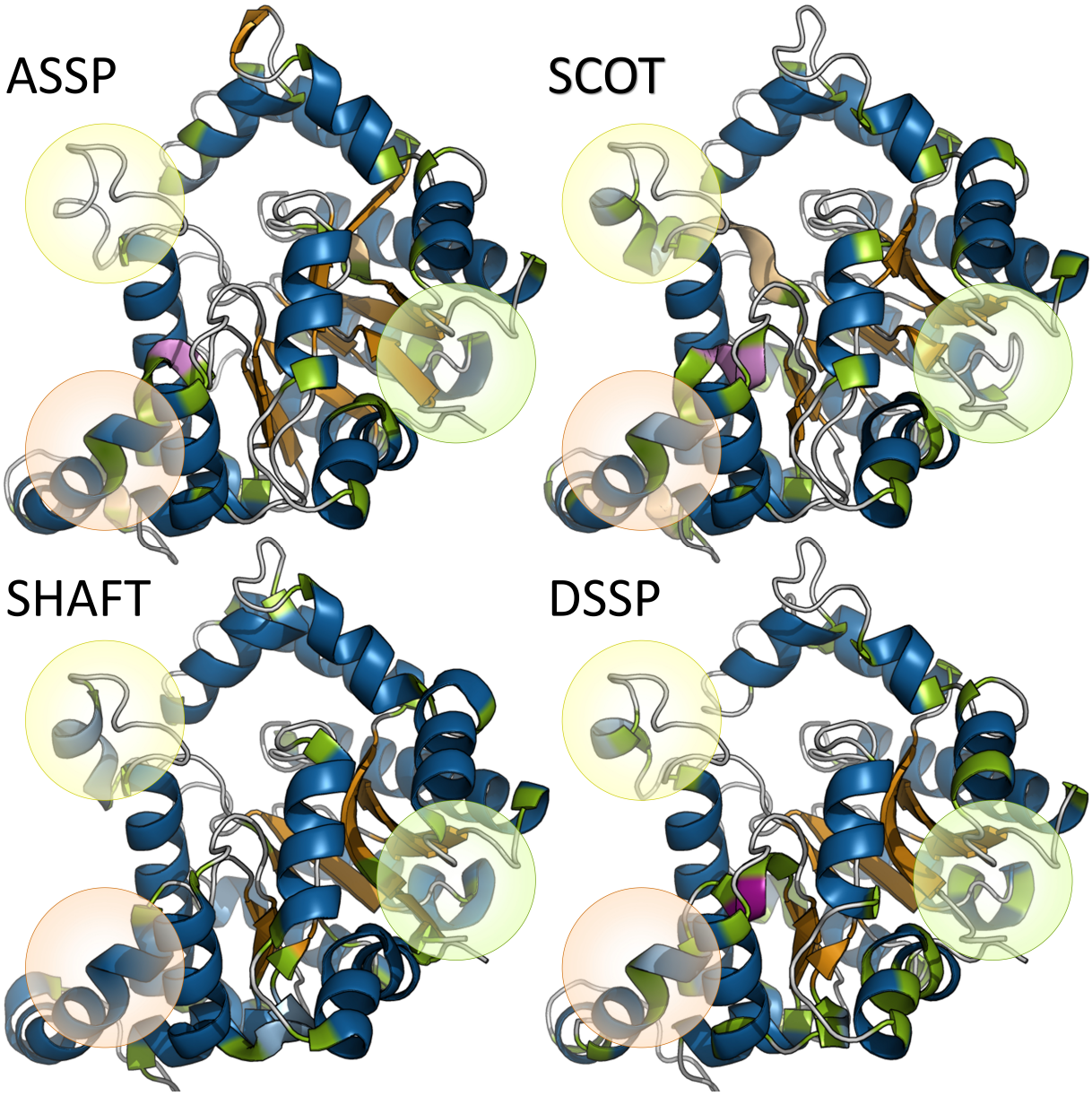


Figure 4.3.20. Comparison of the SSE assignments by different SSAMs for 5-carboxyvanillate decarboxylase. The assignments of right-handed (blue), left-handed (magenta), and PPII (light orange) helices, as well as β-strands (orange) are shown. The helix termini are additionally highlighted in green. The red circles mark 3_{10}-helical backbone regions which are differently classified. An example of a helix with geometric irregularities is highlighted by the red circles, while the green circles focus on an α-helical region including an n6-2 turn with an *i* to *i+5* hydrogen bond.

4.3.4 Conclusion

The novel classification of hydrogen-bonded and open turns in proteins enables the in-depth analysis of various turn types in proteins. Their impact on regular SSEs in proteins was evaluated. Although we find correspondences to Secbase-assigned turn types, major differences highlight the strengths of the novel clustering of turn types based on larger ESOMs. In contrast to the turn assignment procedure implemented in Secbase, the novel turn types offer a better differentiation between open turn types. Furthermore, previously unassigned turn types were classified. The hydrogen-bonded turns are assigned by a more restrictive hydrogen bond criterion leading to a more reliable classification of hydrogen-bonded turns.

The analysis of the relationship between the new turn types and regular SSEs in protein structures enabled the development of a new secondary structure assignment approach, SCOT, which can be readily used as alternative to widely established methods such as DSSP. Apart from that, the new turn classification facilitates the identification of rare and largely unexplored SSEs.

With respect to secondary structure assignment, a method which fulfils all user requirements will probably never exist. The classification of helices and sheets is feasible by visual inspection, by means of hydrogen bonding patterns, or based on geometric criteria. Hydrogen bond-based methods will often assign geometrically distorted helices and kinked β-strands, whereas geometry-based methods, which ignore the presence of intramolecular interactions, even lead to reasonable assignments in the presence of Cα atom coordinates only. Unfortunately, the latter methods are often more restrictive and sensitive towards structure quality and protein flexibility leading to inconsistent assignments. The elaborate combination of both methodologies in SCOT led to optimum results and helped to overcome the major drawbacks of hydrogen bond- and geometry-based approaches.

The impact of secondary structure definition always depends on the final applicability domain. While the development of a secondary structure prediction method highly relies on the accuracy of the underlying training set assignments, geometrically irregular SSEs assigned to assist in a structural alignment might influence the alignment, e.g., with vector-based methods such as LOCK2.[301] Highly bent helices and strands cannot be accurately represented by vectors. In contrast to SHAFT, SCOT enables the consistent classification of geometrically highly regular backbone conformations in proteins.

The analyses also hint at some open challenges for the definition of regular SSEs. The classification of turn types can make considerable contributions to further in-depth investigations on the occurrence of π- and left-handed helices, the assignment of the PPII conformation, and the capping of helices by a multitude of already known and yet to find capping motifs. SCOT was developed to address some of these challenges and offers a novel multi-purpose alternative to the widely applied method DSSP.

As discussed herein and in the doctoral thesis of Dipl.-Inf. Tobias Brinkjost, the incorporation of hydrogen bond and geometric information in SCOT leads to a highly comprehensive assignment software which was shown to be mostly independent of structure quality and flexibility. Its suitability for the SSE-based structure alignment was shown. This latter finding is of importance in the next section when it comes to the identification of SSE-based structural similarities between protein kinases and TryS.

4.4 A Case Study: Trypanothione Synthetase

4.4.1 Motivation

The aim of this case study was to apply acquired knowledge related to protein binding sites and the comprehensively evaluated methods to an interesting target. Different approaches were utilised to enable the structure-based hit identification for a poorly investigated and hardly structurally characterised enzyme. The enzyme TryS was selected to apply these approaches. TryS was shown to be a promising target for the establishment of new therapeutics for the treatment of neglected diseases, such as leishmaniasis. The survival of the causative protozoan organism (*Leishmania*) is dependent on the presence of TryS which fulfils an essential function in the redox metabolism of the parasite.

4.4.2 Comparative Modelling and MD Simulation Studies

Structural knowledge of TryS is restricted to three X-ray structures of the enzyme from *L. major* (PDB-IDs 2vob, 2vpm, 2vps). These structures are characterised by unresolved loop regions and, most importantly, parts of the highly flexible ATP grasp fold are not modelled. Additionally, the binding mode of the substrates has not yet been experimentally verified. The close relationship between TryS and GspS from *E. coli* enabled the generation of a theoretical model of the enzyme in complex with all substrates of the second reaction step (see Section 1.3.2).[162] The alignment of both enzymes reveals crucial similarities with respect to the overall fold, but especially regarding the synthetase domain (see Figure 1.3.2). In contrast, the amidase domain is differently located in both structures. The sequence identity of the synthetase domains of *Ec*GspS and *Lm*TryS is 33% after their structure-based superposition.

Based on the structures of *Lm*TryS with the completely modelled loop segments and ATP grasp fold, comparative models were also generated for TryS from *C. fasciculata*, *L. infantum*, *T. brucei brucei*, and *T. cruzi*. All resulting models were subsequently verified using MD simulation studies as outlined in the following section.

First, MD simulations of *Lm*TryS and *Li*TryS were performed. Simulation data for the apo states, as well as the holo states in the presence of ATP or all substrates of the second reaction step, were analysed. The most intriguing finding was the high flexibility of the amidase domain throughout all simulations (see Figures A4.2 and A4.3 in Appendix 4). A potential explanation is the recombinant expression of the enzyme for crystallization and biochemical assay purposes. For *Cf*TryS, it was experimentally shown that the enzyme purified from *C. fasciculata* cell lysates exists in the form of a heterodimer with *Cf*GspS.[441] In contrast, a monomeric form of the enzyme was detected for recombinantly expressed TryS.[442] As the native forms of *Li*TryS, *Lm*TryS, and further enzymes from other parasites were never analysed, it cannot be precluded that the expression of the protozoan enzyme in *E. coli* leads to an altered quaternary structure and a subsequent difference in the enzymatic behaviour and dynamics of the recombinant and native enzymes. Kinetic and inhibition studies for the proteins discussed herein were also realised with the recombinant protein. These analyses led to consistent results for the synthetase activity of the enzyme.[163] It was also shown that the synthetase activity is unaffected by the truncation of the amidase domain.[443]

Therefore, the following analyses were restricted to the enzymes' synthetase domain although all simulations were performed with the monomeric structure composed of both domains (in accordance with the experimentally used constructs).

The results of the analysis of the atomic fluctuations in the simulations of the *Lm*TryS and *Li*TryS structures can be found in Figure 4.4.1 (see Figures A4.4 and A4.5 for RMSD plots and structural snapshots from the clustering analyses). Overall, we observe a good agreement between the backbone fluctuations of all simulations and the Cα temperature factors of the crystal structure of *Lm*TryS with the PDB-ID 2vpm. Unresolved structural parts of the *Lm*TryS and *Li*TryS structures show high RMSF values, whereas most of the residues involved in GSH and GSP binding are characterised by lower flexibility. In contrast, the backbone of residues involved in ATP binding is less stable. For the simulations of *Lm*TryS in the presence of ATP or of all substrates, some residues of the ATP grasp fold are more stable, while the *Li*TryS simulation in the presence of ATP shows higher flexibility in this part of the nucleotide binding site. The loop region connecting the ATP grasp (referred to as "hinge region" herein) with the synthetase domain shows the highest flexibility in the apo states. However, no global movement of the lid region closing the ATP binding site was monitored.

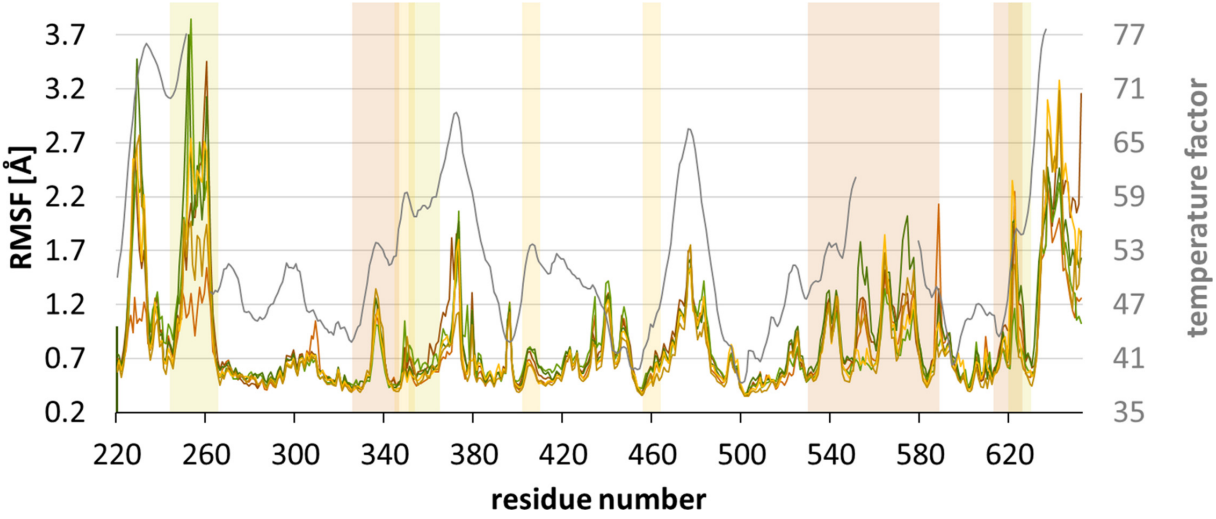


Figure 4.4.1. RMSF plots for the synthetase domain (residues 220 to 652) of *Lm*TryS and *Li*TryS. The per-residue Cα RMSF values for the 600 ns MD simulations of *Lm*TryS (red), *Li*TryS (dark red), *Lm*TryS and *Li*TryS in complex with ATP (green and dark green) and all substrates (yellow and dark yellow) are presented. Residues in the proximity of the ATP, GSH, and GSP binding sites are highlighted by red, yellow, and green bars, respectively. The temperature factor of the Cα atoms of chain A of the structure with the PDB-ID 2vpm is given in grey.

TryS from *T. brucei brucei* (UniProt accession: Q586P2) and *T. cruzi* (UniProt accession: Q9GT49) were also modelled and MD simulations were performed. Selected outcomes will be discussed in the following section. Although the sequence identities of *Tc*TryS and *Tb*TryS to *Lm*TryS are only 58.5% and 58.2% (the sequence identity between *Lm*TryS and *Li*TryS is 95.2%), comparative modelling studies using the *Lm*TryS structure with the modelled grasp fold were performed. The MD simulations reveal a stable structure with a behaviour that was already observed for *Lm*TryS and *Li*TryS (Figure 4.4.2, Figure A4.6 in Appendix 4).

For *Tb*TryS, there is a difference to earlier modelling studies which used the sequence of the UniProt entry with the accession number Q82MC2.[170] This sequence does not correspond to the recombinant protein which Maiwald and co-workers[170] used for enzymatic testing. The plasmid for the expression of the enzyme was obtained as a gift from A. Fairlamb and corresponds to another published sequence (Q586P2).[164] Therefore, only results for this construct are presented as it was used for the inhibition studies cited herein.[168,444]

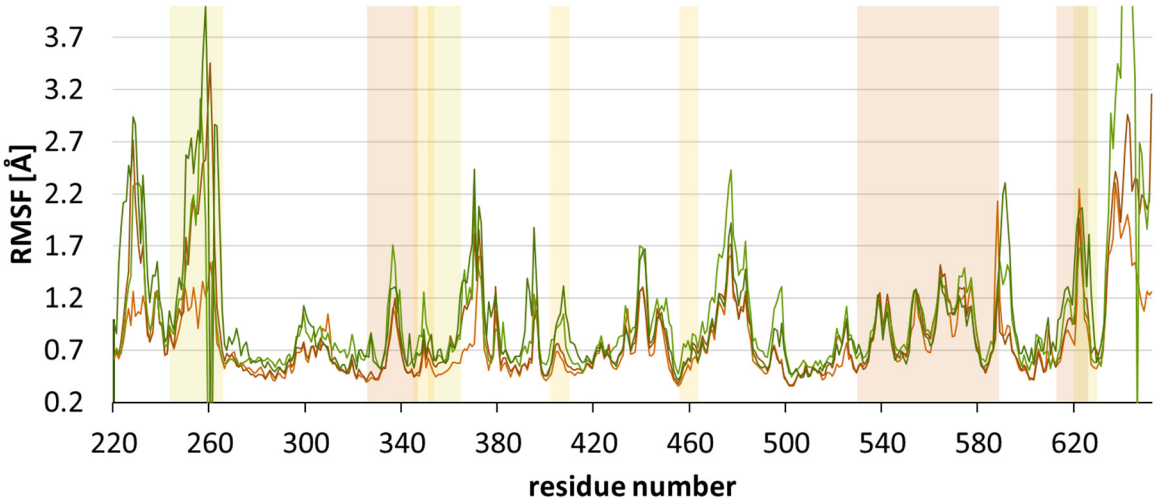


Figure 4.4.2. RMSF plots for the synthetase domain of *Tc*TryS, *Tb*TryS, *Lm*TryS, and *Li*TryS (residues 220 to 652) in the apo state. The per-residue Cα RMSF values for the 600 ns MD simulations of *Lm*TryS (red), *Li*TryS (dark red), *Tc*TryS (green), and *Tb*TryS (dark green). Residues in the proximity of the ATP, GSH, and GSP binding sites are highlighted by red, yellow, and green bars, respectively.

4.4.3 Druggability Assessment

The DoGSiteScorer[254] method was applied to assess the druggability of the ATP binding sites in *Lm*TryS and *Li*TryS. All representative structures derived from the MD simulations in the presence of ATP were used to circumvent inconsistencies in the binding site descriptor calculation arising due to the impact of binding site flexibility (see Section 4.2.5 for the discussion concerning the high variations of the druggability scores for NMR ensemble structures). The ATP molecule was used as binding site-defining ligand. The druggability scores (one is based on a trained SVM and another on a simple regression model) range between 0 for non-druggable and 1 for druggable sites.

For the four *Li*TryS representatives, the mean SVM-based druggability score of 1 hints at the druggability of the site under investigation. The regression-based simple score was also 1 for all representative site conformations. This trend slightly changes for the nine *Lm*TryS-derived sites. However, the mean SVM-based score of 0.96 and the mean regression-based druggability score of 0.97 show that the sites of both enzymes can be addressed by a drug-like small molecule (based on the DoGSite benchmark studies, the optimum threshold for the SVM-based druggability score was defined as 0.5).

These results in hand, suitable ways of addressing the ATP binding site of TryS with a competitive small molecule inhibitor were explored.

4.4.4 The Paullone Binding Mode

The inhibition by compounds with paullone scaffolds is regarded as the link between the TryS and protein kinase ATP binding sites. Paullones in both sites are thought to show a common hinge binding interaction.[444] For some compound classes bearing a paullone scaffold and inhibiting trypanosomal growth, TryS was already excluded as a potential target.[170] Instead, kinases of the parasites are probably addressed by some paullone-based compounds. Up to date, crystal structures of TryS from organisms of the genus Leishmania in complex with the most potent inhibitors MOL2008 or FS-554[168] are not publicly available. Both compounds show a high selectivity towards *Li*TryS as compared to the trypanosomal enzymes.

Docking studies and subsequent MD simulations were applied to obtain a reasonable hypothesis for a potential binding mode. The initially suggested binding mode for compound MOL2008[444] was found as not being stable in the corresponding MD simulation for *Li*TryS (see Figure A4.7 in Appendix 4 for the RMSD plots of the simulations of *Li*TryS with the inhibitor MOL2008 in the proposed binding mode and snapshots of the initial binding mode and the binding mode observed after 600 ns simulation time). The binding mode of the compound in the final equilibrated complex is highly solvent-exposed. The interaction pattern does not correspond to the initial binding hypothesis which was derived based on the similarity to the protein kinase GSK-3β. Given this finding, an independent prediction of a probable binding mode for N^5-substituted paullones in TryS from Leishmania was pursued.

Several facts supported the search for probable binding modes (see Figure A4.1 for the published bioactivity data). First, the most potent inhibitors show low IC_{50} values for *Li*TryS, while the inhibition of TryS from *T. brucei brucei* and *T. cruzi* is negligible. The omission of the N^5-substitution in both *Li*TryS inhibitors leads to significantly lower potency. Consequently, a spermidine competitive mechanism of inhibition was proposed.[168] The absence of the terminal secondary amine at the N^5-substitution leads to a higher inhibitory potential against *Tc*Trys, whereas higher residual activities at a compound concentration of 30 µM are observed for *Li*TryS. A paullone scaffold-based inhibitor with high selectivity and inhibitory potency for *Tb*TryS could not be identified in the studies of Orban *et al.* although the EC_{50} values for *T. brucei brucei* are low, indicating a different primary target in Trypanosoma.[444] All these findings are in line with the new results which will be summarised in the following paragraphs.

After the molecular docking of FS-554 and subsequent MD simulations of the theoretically proposed complex structures, only one binding mode was found to be stable throughout the MD simulations with *Li*TryS as well as *Lm*TryS. The binding mode is presented in Figure 4.4.3 and does not correspond to the experimentally verified binding mode of the alsterpaullone to GSK-3β. The paullone scaffold in the protein kinase structure with the PDB-ID 1q3w (chain A) is engaged in hydrogen bond interactions with the main-chain of Val315 in the hinge region. In contrast, we observe main-chain hydrogen bonds to the backbone nitrogen atom of Phe586 and the carbonyl oxygen atom of Val618 in TryS from Leishmania. An intramolecular hydrogen bond stabilises the conformation of FS-554 in the front part of the ATP binding site (see Table A4.3 for the hydrogen bond analysis of all MD snapshots).

Phe343 is involved in a π-stacking interaction, while the trifluoromethyl substituent is located in a hydrophobic sub-pocket which is separated from the solvent by the side-chain of Arg613.

Most importantly, we observe a hydrogen bond interaction between the secondary amine and Glu587. Sequence alignments of TryS from *L. infantum*, *L. major*, *T. cruzi*, and *T. brucei brucei* show a substitution at this position (Table A4.4), whereas the other residues that interact with FS-554 are identical. TryS from different strains of Trypanosoma harbour an Arg instead of a Glu. This residue forms a stable hydrogen bond to Asp322 of *Tc*TryS and Asp328 of *Tb*TryS (see Table A4.2 in Appendix 4 for details on the hydrogen bond interactions of this Arg residue during the MD simulations). Consequently, the negative charge is missing in the binding sites of these enzymes. This is a plausible explanation for the missing inhibitory potential of FS-554 on TryS from both organisms. The replacement of the amine by an aromatic substituent and the resulting increase in inhibitory potential is in line with the proposed binding mode. The presence of hydrophobic residues in the trypanosomal enzymes might explain this trend.

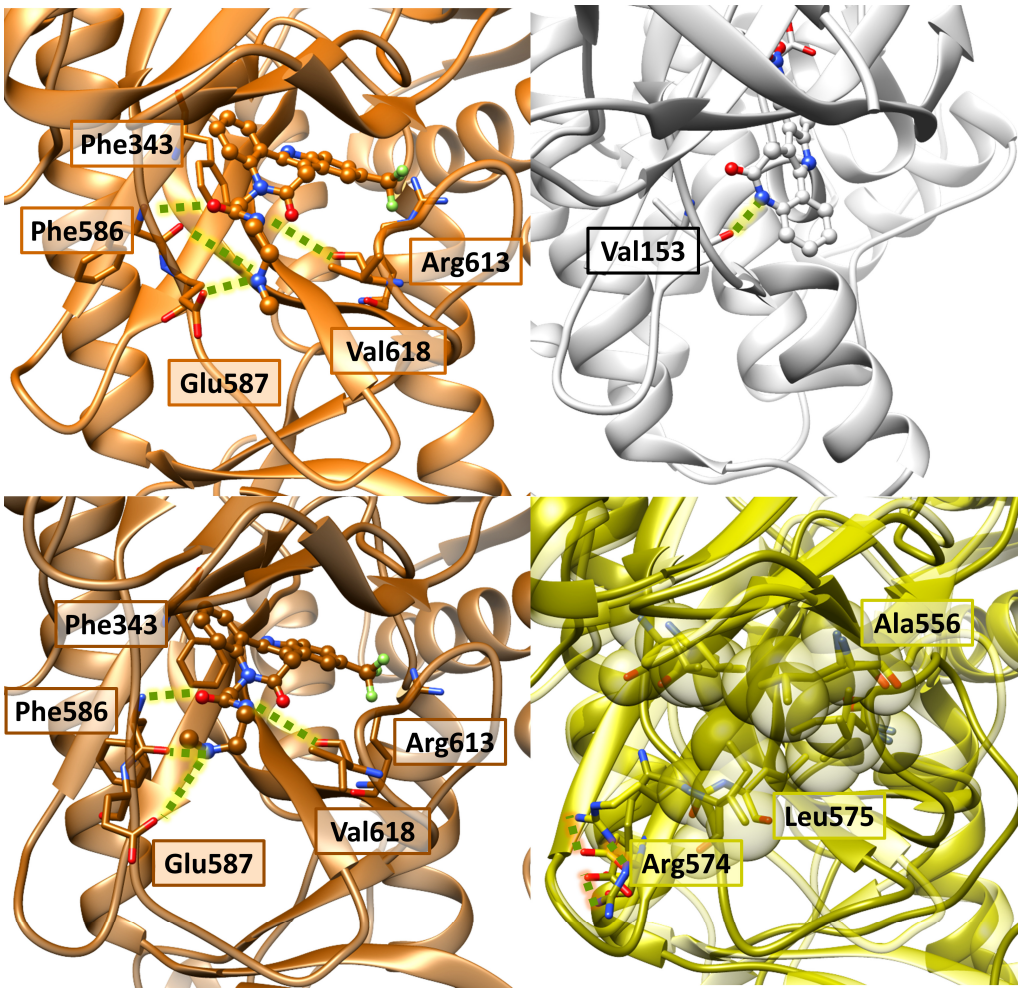


Figure 4.4.3. Proposed binding mode of *Li*TryS inhibitor FS-554. The final snapshot of the MD simulation for the proposed FS-554 binding mode in *Lm*TryS (top, left) and *Li*TryS (bottom, left) is given. All presented TryS structures are aligned with respect to the Cα atoms. The hydrogen bonds of FS-554 (ball-and-stick) to the protein are highlighted by dotted lines. The binding mode of the alsterpaullone in GSK-3β (PDB-ID 1q3w, chain A) after manually aligning the TryS and protein kinase ATP binding sites is shown in the top right panel. Final MD snapshots of simulations of *Tb*TryS and *Tc*TryS are provided in the bottom right corner. The hydrophobic residues in *Tb*TryS (Leu562, Leu569, Ile582, and Leu612) are additionally represented as spheres to highlight the closed state of the front pocket in this enzyme.

There is one major difference between *Tc*TryS (inhibited by paullones with alkyl- and aryl-modifications at position N^5) and *Tb*TryS (no significant inhibition by paullone-based compounds): Leu588 at the pocket entrance which is found in *Li*TryS, *Lm*TryS, and, *Tc*TryS is substituted by an Ile in *Tb*TryS leading to a complete closure of the front part of the pocket mediated by the hydrophobic residues Leu562 (Ala556 in *Tc*TryS), Leu579, Ile582 (Leu575 in *Tc*TryS), and Leu612 (see Figure 4.4.3, bottom right panel). This precludes the observed binding mode for *Tb*TryS and might also explain why trimming down the N^5-substituent leads to more potent *Tc*TryS and *Tb*TryS inhibitors.[445]

After all, a novel binding hypothesis of the paullone FS-554 in TryS could be derived. Intriguingly, this binding mode is significantly different from the alsterpaullone binding mode in GSK-3β. Therefore, paullone-inhibited protein kinases (e.g., PHK, GSK-3β, CDK2 (cyclin-dependent kinase 2), PAK4, PAK5)[446] might not be appropriate starting points for drug repurposing strategies. Despite these investigations, links between ATP grasp enzymes and protein kinases proved successful in the identification of novel ATP competitive inhibitors.[167] Therefore, the following paragraphs deal with investigations concerning similarities between TryS and protein kinases.

4.4.5 The Initial Search for TryS Inhibitors

Focusing on the previously identified similarities between TryS and protein kinases, a molecular docking workflow was applied to screen the Kinase SARfari database[447] for interesting kinase inhibitors with potential bioactivities on TryS from *L. major*. This theoretical study was based on MD-derived structural representatives of *Lm*TryS. Representative structures of the ATP binding site were extracted for an ensemble docking approach. The established workflow as pursued by Nana Boateng during his bachelor studies[448] consisted of constraint-based molecular docking, score normalisation, filtering steps, and a subsequent visual inspection of all molecules which showed binding modes with high scores. The results not only led to interesting molecules for further testing, but also indicated similar kinases whose inhibitors were enriched within the best scored molecules (including JAK2, CHUK, MAP3K5, ROCK2, and NF-κB). However, the exclusion of inhibitors with known bioactivities on more than three kinases might have biased these results.

The commercial availability of the finally chosen compounds restricted the number of small molecules to be tested. Consequently, only four potential ATP pocket binders were tested in the laboratory of M. Comini in Uruguay. The compound structures and their corresponding docking modes in *Lm*TryS are given in Figure 4.4.4. A binding motif which was found for all four compounds consists of two hydrogen bonds towards the backbone atoms of residues Gln584 and Phe586. These interactions were initially supposed to be crucial for TryS binding and set as protein hydrogen bond constraints in a post-processing step of the molecular docking. For the compounds with the ChEMBL-IDs 383482 and 504950, one further hydrogen bond to the side-chain of residue Gln583 was observed in the resulting docking poses. Additionally, all but one compound were predicted to interact with Lys548 via a hydrogen bond.

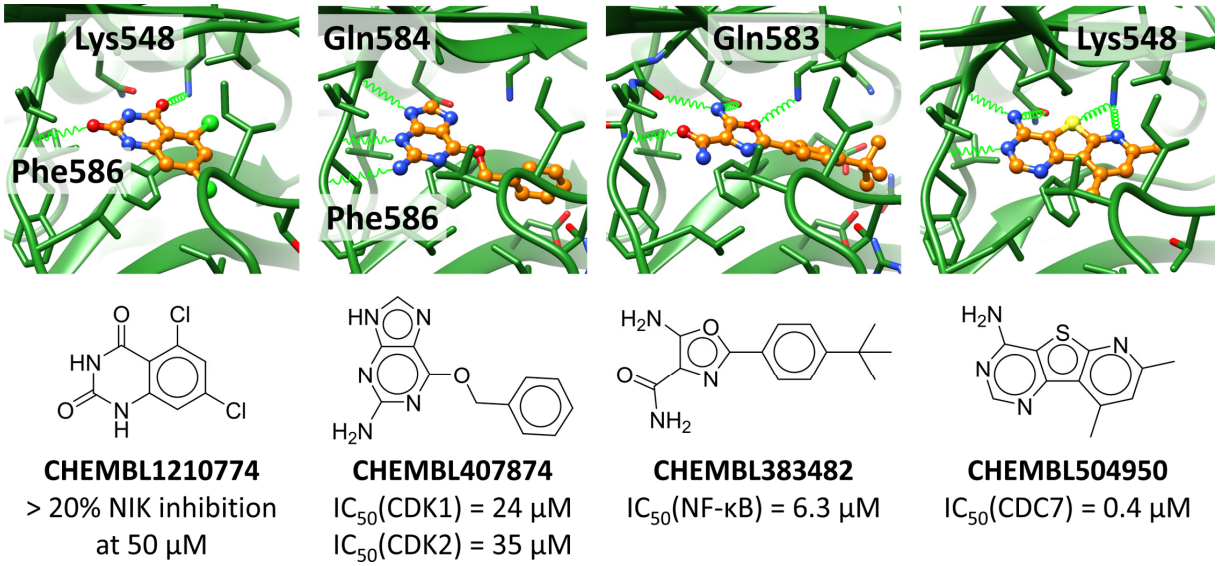


Figure 4.4.4. Molecular docking poses in *Lm*TryS, structures, and known targets of four small molecule kinase inhibitors. The compounds were purchased and biochemically tested based on the molecular docking results of the Kinase SARfari database. Their inhibitory potential on kinases is given below the ChEMBL-ID. The hydrogen bond interactions of all compounds (orange, ball-and-stick) are highlighted by green springs.

The biochemical results for TryS from three different, therapeutically interesting organisms are provided in Table 4.4.1. TryS from *L. major* was not tested, but the results for *Li*TryS should predict the probable outcomes for *Lm*TryS due to the highly similar ATP binding sites and the high overall sequence similarity. CHEMBL504950 shows at least some inhibitory activity on *Li*TryS. Although the residual activity of some enzymes could be decreased to a minor extent by at least one of the compounds, the outcome of this first analysis did not lead to significant hits worth further investigations. The question arises whether the elucidation of highly similar kinases to generate a focused kinase inhibitor library might result in more promising results. This approach is discussed in more detail in the following two sections.

Table 4.4.1. Biochemical assay results of four hit compounds as identified by the molecular docking of the Kinase SARfari database. The per cent residual activity of *Tc*TryS, *Tb*TryS, and *Li*TryS at a final compound concentration of 30 µM is given together with the corresponding standard deviations.

	***Tc*TryS Activity [%]**	***Number of Repetitions***	***Tb*TryS Activity [%]**	***Number of Repetitions***	***Li*TryS Activity [%]**	***Number of Repetitions***
CHEMBL1210774	82.1 ± 3.4	*3*	90.9 ± 9.7	*4*	93.7 ± 4.6	*2*
CHEMBL407874	94.9 ± 4.9	*4*	92.4 ± 4.5	*4*	88.6 ± 9.6	*4*
CHEMBL383482	92.2 ± 6.7	*3*	85.5 ± 6.9	*4*	102.0 ± 8.3	*3*
CHEMBL504950	94.3 ± 8.4	*4*	85.2 ± 6.0	*4*	91.3 ± 3.6	*4*

4.4.5 TryS and Protein Kinases – A Common Ligand-Sensing Core?

The search for similarities between TryS from two species of Leishmania and protein kinases was initially performed based on the ligand-sensing cores concept.[68] A similar arrangement of SSEs in the immediate proximity of the ATP sites in both TryS and certain protein kinase families might assist in the identification of structurally related protein kinases enabling a focused search for novel potential ATP pocket binders.

To this end, the MD-derived representative ATP binding sites of *Li*TryS and *Lm*TryS were compared to ligand-bound main pockets of the protein kinases from the KLIFS database.[302] The comparison of the SSE arrangments was realised by two methods. First, the graph-based approach SLOT was used. This method was developed in cooperation with Dipl.-Inf. Tobias Brinkjost who implemented SLOT as a standalone software which incorporates the structure preparation, the site extraction, the graph modelling, the graph-based comparison, and the export of the resulting alignments. Second, the vector-based algorithm LOCK2 was applied.

The alignment of GSK-3β and TryS (see Section 1.3.3) was most promising for SLOT. The strand arrangements that are found in nearly all protein kinases and the grasp fold of TryS were matched, but also an additional helical structure. This match was used as a test case in all alignment studies as it is the only protein kinase structure in complex with a paullone scaffold. The SLOT parameters were optimised with respect to this match which led to highly non-restrictive matching parameters. To exclude the possibility that these optimisation steps lead to a too permissive matching and unusual results for alignments to other protein kinase structures, LOCK2,[301] as an alternative SSE matching approach, was applied to compare TryS to the protein kinases. It aligned GSK-3β and TryS as expected, but the number of matching SSEs was significantly lower than that obtained by SLOT (Figure 4.4.5).

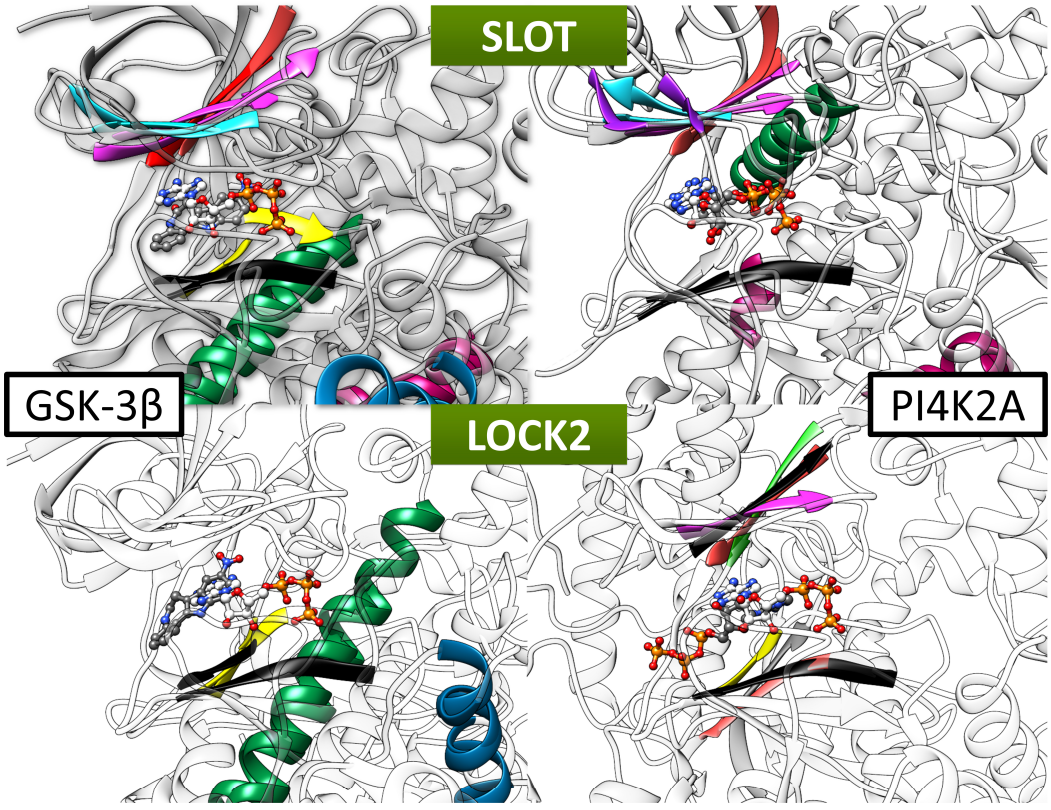


Figure 4.4.5. Alignments of TryS from Leishmania in complex with ATP and protein kinases. The SLOT-derived alignments of *Li*TryS to the protein kinases GSK-3β (PDB-ID 1q3w, chain A) and PI4K2A (phosphatidylinositol 4 kinase type IIα, PDB-ID 4hnd, chain A, in complex with ADP) are given at the top. The LOCK2-derived matches of *Lm*TryS to GSK-3β (PDB-ID 1q3w, chain A) and PI4K2A (PDB-ID 4pla, chain A, in complex with ATP) are shown at the bottom of the figure. The TryS structures are shown in white, while the protein kinases are shown in grey. The matched SSEs are coloured according to the SLOT-generated reference match (top, left, shaded). SSEs which were not matched in the reference alignment were assigned new colours.

Encouraged by the successful superposition of TryS and GSK-3β, the complete synthetase domain of TryS (query structures) was compared against all protein kinase structures in the KLIFS database (target structures). In a second step, excised binding sites of both query and target structures were used as input. As depicted in Figure 4.4.6 and Figure A4.10 in Appendix 4, it is difficult, if not impossible, to pinpoint a single kinase, kinase family, or kinase group which is most similar to TryS based on similarity scores and further characteristics of the resulting matches.

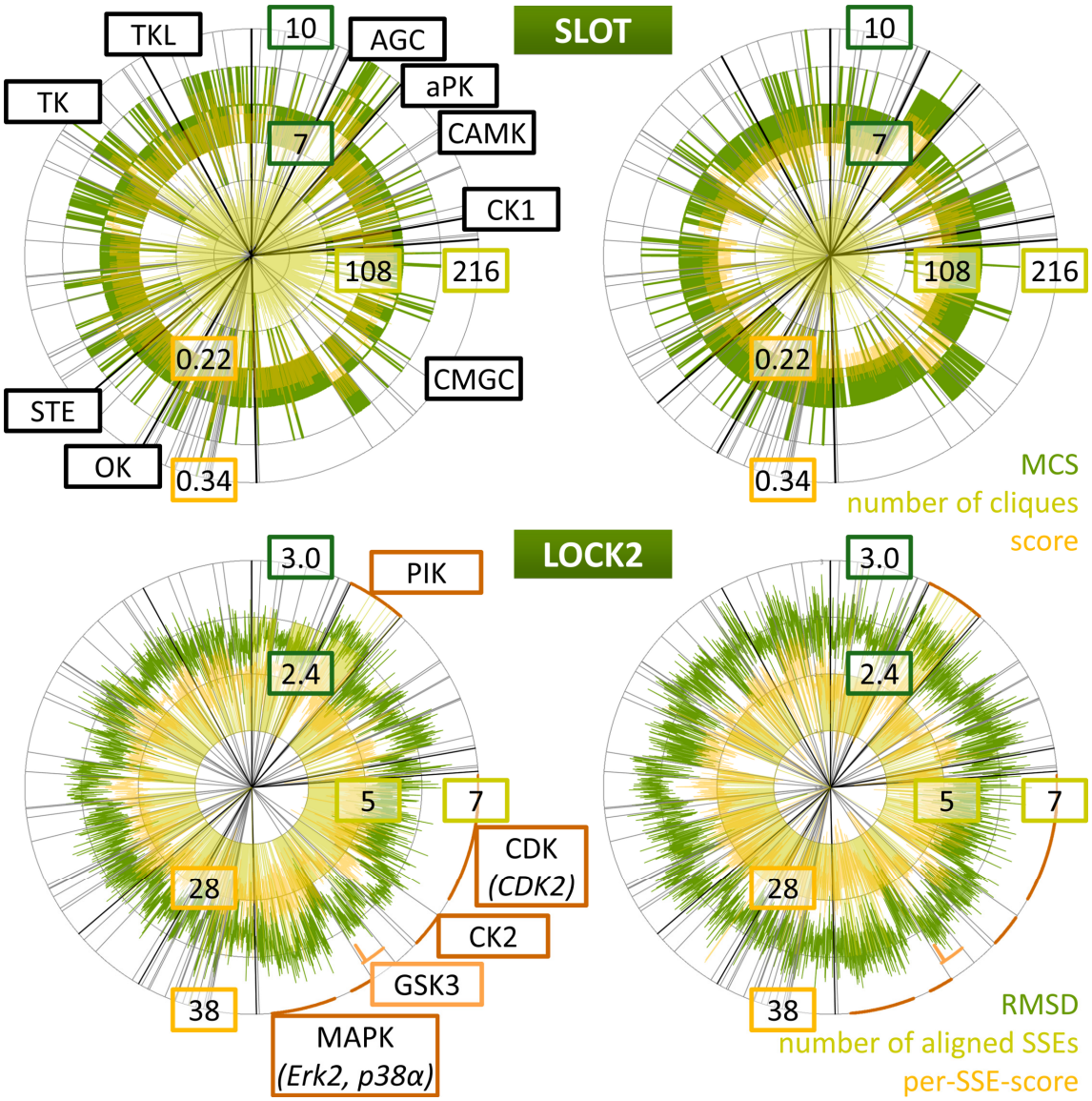


Figure 4.4.6. Secondary structure similarities between TryS and protein kinases. The radar plots of MCS (number of matched SSEs), number of cliques, and score for the SLOT-derived matches (top) and RMSD, number of aligned SSEs, and per-SSE-scores for the LOCK2 alignments (bottom) of TryS structures (synthetase domain) with protein kinase structures of the KLIFS database are given. The results for the comparison between *Lm*TryS and *Li*TryS representative structures are given on the left and right, respectively. The kinase structures are grouped according to their membership to protein kinase groups as defined by Manning and co-workers[449]: AGC (protein kinase families A, G, and C), aPK (atypical protein kinases), CAMK (Ca^{2+}/calmodulin-dependent protein kinase), CK1 (casein kinase 1), CMGC (group including CDK, MAPK (mitogen-activated protein kinase), GSK, and CDK-like kinase families), OK (other kinases including e.g., Aurora kinases and CDC7), TK (tyrosine kinase), and TKL (tyrosine kinase-like).

Although both methods, SLOT and LOCK2, find the anticipated match between TryS and GSK-3β in complex with the alsterpaullone, the LOCK2-based alignments of the TryS structures to other kinases do not lead to meaningful results in most cases. Either the ATP binding sites do not overlay as expected or the quality of the SSE alignment is low.

Nevertheless, kinase families with on average lower RMSD values, higher per-SSE-scores, or higher numbers of matched SSEs were identified. The families phosphoinositide kinases (showing a high per-SSE-score and higher numbers of matched SSEs), CDK2 structures (showing a low RMSD for most matches), the casein kinase 2 family (high per-SSE-score), and Erk2 and p38α structures (low RMSD) were analysed in more detail. In all cases, the alignments were not as expected as the ATP molecules of the binding site could not be properly aligned (see the right panel in Figure 4.4.5 for an example). Furthermore, different representative TryS structures led to highly different alignments underlining the missing robustness.

In contrast, the alignments using SLOT often represent the anticipated structural overlays with the kinase ligands and the ATP molecule in the TryS structure. When comparing complete chains, we observe a high number of matched SSEs in TryS to those of structures of the PIK family (atypical kinase group) for the SLOT comparison method. However, all structural matches are characterised by a high number of possible alignments (i.e., a high number of cliques) and it is not possible to further distinguish the different superpositions with respect to their quality. This can be attributed to the highly permissive parameters which are necessary to match the SSE arrangements of TryS and kinases. The increase in the number of matched SSEs is explicable when considering the structures of PIK family members. The structures are larger than those of other protein kinases and the number of helices and strands to be matched is consequently higher.

Another weak point of the approach is the nature of the alignments. The main SSE-based similarities between kinase and TryS structures result from a similar arrangement of six β-strands surrounding the binding site. This leads to the commonly observed MCS size (number of matched SSEs) of seven to nine and an equally distributed score across all kinase families. Based on the alignment characteristics of the LOCK2-based alignments (RMSD, per-SSE-score, and the number of aligned SSEs), the test case match with GSK-3β is not distinguishable from all other TryS-kinase matches. Similar ranges for these similarity metrics are found for all pairwise comparisons. In contrast, the number of matched SSEs between protein kinases of the GSK family and *Lm*TryS is higher using SLOT. This finding points towards a potential bias of the parameters which were optimised for these matches. The lower number of matched SSEs between *Li*TryS and GSK kinases further substantiates this conclusion.

For pocket comparisons with LOCK2, the results were even less distinctive and no kinases with significantly higher similarities to the ATP sites of *Li*TryS and *Lm*TryS could be identified (Figure A4.10). In contrast, the SLOT comparisons reveal a potential relationship between TryS and the kinase PIM1. One or two more SSEs are matched in comparison to other protein kinases. However, the number of cliques is as high as for all other kinases with a maximum of 61 possible alignments of one kinase structure to *Lm*TryS, hampering reliable conclusions.

In summary, an application of the ligand-sensing cores approach for the identification of the most similar protein kinases with respect to TryS is questionable. Although the SSE arrangements of a small set of protein kinase families seem to be most similar to those of the ATP grasp enzyme, a reliable differentiation based on the pairwise alignments is impossible. This was shown for two independent methods developed for the automated comparison of SSE arrangements in proteins. The SSE-based similarities are not sufficient to name an interesting kinase family for drug repurposing studies. Consequently, an alternative and commonly accepted approach was utilised to search for physicochemically similar protein kinase binding sites. The respective outcome is outlined in the following section.

4.4.6 TryS and Protein Kinases – Similarities in the ATP Binding Site Properties

The most promising binding site comparison software (see Section 4.2.8) was used to evaluate potential similarities between the TryS and protein kinase ATP binding sites. KRIPO was not used as the tool did not offer the possibility to create fingerprints for user-defined protein structures at the time of these analyses. A consensus approach for binding site comparison was established as the results in Section 4.2.5 hint at the complementarity of different binding site comparison tools.

The protein kinase structures of the KLIFS database were used as target structures for pairwise binding site comparisons to the MD-derived representative structures of *Li*TryS and *Lm*TryS. All known kinase structures were used instead of using only one structure per protein kinase to circumvent problems arising from the sensitivity of some binding site comparison methods with respect to the binding site definition and flexibility (see Section 4.2.5 for details).

A filtering step was performed to post-process the obtained results. Given the results of the binding site comparison benchmark studies, matches with considerably low scores were excluded. To this end, only matches with scores above the upper whisker of the boxplots for the dissimilar pairs were considered (see Figures A2.3-5 in Appendix 2). This restriction to site pairs with high similarity scores led to exceedingly small numbers of matches for FuzCav, PocketMatch, and RAPMAD. Considering this and their inability to provide the corresponding site alignments (see Section 4.2), their results were not considered in the following analyses.

The comparably good performance of all remaining comparison methods hampers the decision for one specific tool. Instead, a consensus approach, which already proved useful for other structure-based methods, was applied. All identified similar binding site pairs were re-evaluated by the number of methods that found the specific site pairs. Out of the nine finally applied tools, a maximum consensus of seven was yielded for the comparison to the nine *Lm*TryS representatives. For *Li*TryS, the number of representatives was lower than for *Lm*TryS (four structures) which is also reflected by the lower maximum consensus of six tools and the lower number of detected similarities (Tables A4.5 and A4.6).

KinMap[450] was initially used to plot the matches and get an overview of the diversity of the protein kinases whose ATP binding sites were often identified as similar to the TryS binding sites (see Figure A4.11 in Appendix 4).

At first glance, there does not seem to be a significant trend towards a special subset of kinases. For CDK2 and GSK-3β – the kinases which were expected to show the highest similarities to TryS as they are inhibited by paullone-based compounds – the number of matches with the representative *Li*TryS binding sites was low as compared to other kinases. In contrast, *Lm*TryS structures are more often matched to both kinases. A distinct trend which could be seen for both *Li*TryS and *Lm*TryS was found with respect to the bound ligands of the matched protein kinase binding sites. Matches towards kinases in complex with ATP, ADP, and their analogues (with the IDs ANP, AGS, etc.) were enriched within the filtered results. This finding underlines the significant impact of bound ligands on the comparison. However, the high number of similar protein kinases – although not surprising – still hampers the decision for specific kinases to analyse the suitability of a drug repurposing strategy. Therefore, an additional filtering criterion was applied to find the most promising kinase matches.

Of all analysed tools in the binding site comparison benchmark studies, SiteHopper proved outstandingly successful regarding the ranking of binding site pairs with increasing dissimilarity. Consequently, the SiteHopper scores for all filtered consensus matches were analysed to elucidate the most similar kinases (see Tables A4.5 and A4.6 in Appendix 4). The overall PatchScore unifies two measures for similarity: the ShapeTanimoto (shape similarity) and the ColorTanimoto (physicochemical similarity). A purely shape-based similarity will not be sufficient for a successful repurposing strategy. Therefore, we focused on matches with a high degree of common physicochemical features.

The structures of two protein kinases, Erk2 and PKACA, gave the highest PatchScores for matches with *Li*TryS and *Lm*TryS, and showed a sufficiently high ColorTanimoto. This is also reflected in the underlying SiteHopper alignments which are presented in Figure 4.4.7. The binding sites of TryS and both kinases and the corresponding ligands overlap as expected for the ATP binding sites, and similar residues can be highlighted. In contrast, the ColorTanimoto for matches of TryS to the protein kinases DAPK1, MAP2K1, and p38α was significantly lower, although many consensus matches to these kinases were detected.

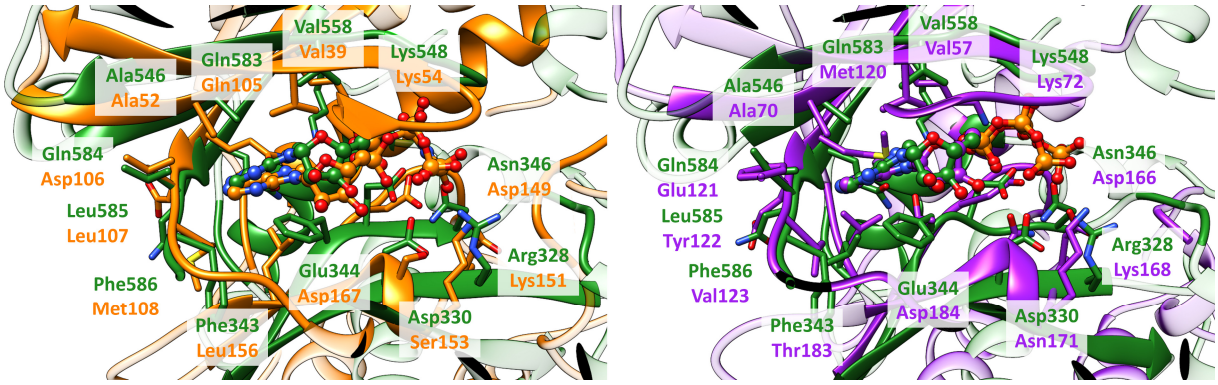


Figure 4.4.7. Alignments of *Li*TryS and the protein kinases Erk2 and PKACA obtained with SiteHopper. On the left, the alignment between the ATP binding sites of Erk2 (orange, PDB-ID 4nif, chain B) and *Li*TryS are presented. On the right, the aligned ATP binding sites of PKACA (purple, PDB-ID 4wb6, chain B) and *Li*TryS are shown. The ligands (*Li*TryS, Erk2, and PKACA in complex with ATP, ANP, and ATP, respectively) are given as ball-and-stick models. Similar residues in both binding sites are represented as sticks and their names and IDs according to the PDB structures are given.

4.4.7 Drug Repurposing and Related Strategies

As outlined before and illustrated by successful hit identification projects,[354,356,360,374–376,451] binding site comparison alone will rarely suffice for a successful drug repositioning strategy. In addition, molecular docking studies, pharmacophore searches, and MD analyses pave the way towards the identification of promising hits. The identification of similar binding sites and the prioritisation of compounds with known affinities towards the most similar pockets can be regarded as a filtering step in VS to reduce the huge chemical space and generate a focused library of potentially active compounds towards the POI. This final section will discuss two potential hit discovery strategies which are based on the binding site comparison results for TryS. The first one is a pure drug repurposing strategy, while the second approach is aimed at the generation of a focused library and the enrichment of potentially bioactive compounds by molecular docking.

Known complex structures of the most similar kinases, Erk2 and PKACA, were searched for in the PDB and the ATP binding site ligands were extracted for preliminary docking studies to validate the applicability of the intended drug repositioning strategy. Most Erk2 inhibitors are characterised by a large hinge-binding element which cannot be appropriately accommodated within the front part of the TryS binding site due to its closed conformation (Figure A4.12 in Appendix 4). Therefore, only compounds with a small hinge-binding element were appropriate for testing purposes. In contrast, most PKACA inhibitors are characterised by a sterically less demanding hinge-binding motif (see Figure 4.4.8 for the proposed binding modes of the kinase inhibitors in TryS). However, they are very small and more elaborate structure-based design studies will be necessary to derive compounds which address the TryS ATP binding site but do not bind to the ATP sites of protein kinases.

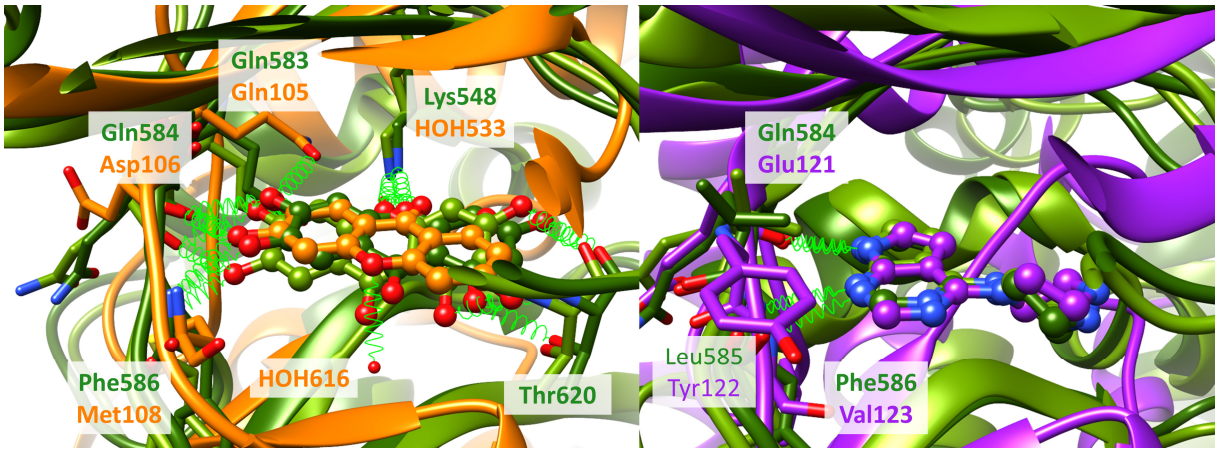


Figure 4.4.8. Proposed binding modes of known protein kinase inhibitors to TryS. Docking-based binding modes of compounds which bind to either Erk2 (left) or PKACA (right) and show similar binding modes in the molecular docking results with representative ATP binding sites of *Li*TryS (dark green) and *Lm*TryS (light green) are presented. The original binding modes of norathyriol in Erk2 (ligand-ID NRA, PDB-ID 3sa0) and an adenosine analogue in PKACA (ligand-ID 6J9, PDB-ID 5izf) and their predicted binding modes in *Li*TryS and *Lm*Trys (ball-and-stick representation) are presented. Hydrogen bond interactions are indicated by green springs and the corresponding residues are shown in stick representation. The structures of the protein kinases and TryS were aligned using SiteHopper.

Inhibitors of both kinases as retrieved from the ChEMBL[303] database were also used for an ensemble docking using representative *Li*TryS and *Lm*TryS structures. However, a visual inspection of the corresponding binding modes showed that this approach did not lead to satisfactory results. The predicted poses were characterised by a small number of interactions and comparably low scores. Therefore, an alternative approach was pursued to search for compounds with a potential affinity towards the ATP binding site of TryS.

A database of commercially available compounds was screened for molecules with similarities to known Erk2 and PKACA inhibitors from the ChEMBL (see Section 3.4.6). The identified compounds were subsequently docked into the binding sites of *Li*TryS and *Lm*TryS.

The visual inspection of the resulting poses revealed a common binding type for the highest scored hits. An example is provided in Figure 4.4.9. Most compounds undergo hydrogen bond interactions with the hinge region and two Lys residues situated the centre of the ATP binding pocket (highlighted in green in Figure 4.4.9). The subpart of the pocket that is originally occupied by the γ-phosphate of ATP is addressed by amide and sulfonamide moieties (highlighted in red), while aromatic and alkyl substituents address the back part of the pocket which is flanked by the hydrophobic residues Val554 and Val622 (highlighted in yellow). The presented compound is large as compared to usually tested hit compounds, but it is used to illustrate all crucial interactions which were also observed for other selected compounds (see Tables A4.7 and A4.8 in Appendix 4 for an overview of their structures). The selected VS hits are representatives of the dataset and for each molecule, many highly similar molecules led to similar proposed binding modes and comparably high scores. Intriguingly, they are all based on scaffolds which were found in the ChEMBL-derived Erk2 inhibitors. The docking-based proposed binding modes of all promising compounds show stable interactions with a similar set of residues in TryS. Intriguingly, most of these interacting residues were found to be similar to the residues in the binding sites of Erk2 and PKACA. Experimental testing of these molecules could pave the way to novel hit compounds for the inhibition of TryS.

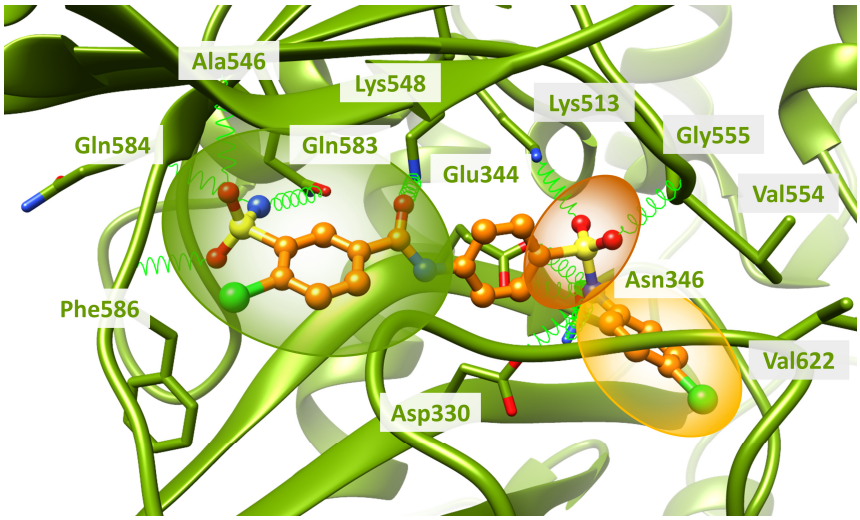


Figure 4.4.9. The binding mode of a promising compound in *Lm*TryS. The compound with the MolPort-ID 005-832-541 (orange, ball-and-stick representation) interacts with *Lm*TryS via many hydrogen bond interactions (green springs). Residues with similar counterparts in the kinases Erk2 and PKACA are given on a white background, while the grey background indicates dissimilar residues.

A drug repurposing strategy failed in the case of both kinases. However, the repurposing of known inhibitors of the kinases PKACA and Erk2 has several drawbacks. The first limitation is highly connected to the problem of kinase inhibitor selectivity, especially in the case of type I inhibitors.[452] The use of known kinase ATP pocket binders to address TryS might harbour potential pitfalls, e.g., off-target effects on human protein kinases. Moreover, the potency range of approved drugs does not necessarily cover the potency range for a new application. Therefore, a pure repositioning strategy might fail. Moreover, the use of PKACA inhibitors for drug repurposing studies is questionable as the enzyme is commonly regarded as anti-target and rarely addressed by drugs.[453] The multitude of cellular processes that are regulated by cAMP underlines the crucial physiological function of this kinase. Finally, the original repurposing workflow was hampered by the need for purchasable compounds. Only 32 of the kinase inhibitors in the ChEMBL[303] database were commercially available and none of them represents an approved drug.

The outcome of the second analysis led to the identification of the most promising compounds for testing. Although they do not represent well-characterised inhibitors of Erk2, they show similarities to known inhibitors (see Tables A4.7 and A4.8 in Appendix 4). Most importantly, their docking-based binding modes in both TryS structures show that they interact with residues which have similar counterparts in the protein kinases (based on the SiteHopper alignment). A comparison of the interacting residues to those in *Tb*TryS and *Tc*TryS additionally shows that the compounds might also inhibit the enzymes from *T. brucei brucei* and *T. cruzi* (see Figure A4.13). Experimentally verified hits in the set of selected compounds might therefore even harbour the potential of being applied as starting points for the development of compounds for the treatment of further diseases that are caused by Trypanosomatida. Due to the enrichment of these virtual hits in the score-ranked lists of the molecular docking and their unique binding modes, the selected compounds are worth testing.

4.4.8 Conclusion

Comparative modelling and MD simulations provided in-depth structural knowledge of TryS from different organisms. The simulation analyses of the structural models in the absence and presence of the substrates underpin the applicability of different modelling approaches based on the initial theoretical structures. Representative snapshots were used for different structure-based approaches to ensure sufficient consideration of protein flexibility in all analyses. The main focus of the analysis was the investigation of the ATP binding site of the enzyme and the search for novel promising compounds which might act ATP competitive, thereby inhibiting TryS. The basic druggability of the cavity was theoretically confirmed and a workflow for the identification of novel hit compounds binding the highly buried ATP grasp fold of TryS was established.

It was previously assumed that the ATP binding sites of *Lm*TryS and the protein kinases CDK2 and GSK-3β share crucial similarities as all these enzymes are inhibited by compounds characterised by a paullone scaffold. However, the nitrogen and oxygen atoms of this scaffold are engaged in hydrogen bond interactions to the hinge region of protein kinases. In contrast, the most promising TryS inhibitors are characterised by an alkylation of the respective nitrogen atom. As their binding mode has never been experimentally verified and it is not known which of the putative binding sites are addressed, a potential binding mode which fits hitherto existing knowledge concerning this compound class was predicted. Although the finally proposed binding mode is in line with the experimental observations, it does not correspond to the respective binding mode of the paullone in the known complex structure of the protein kinase GSK-3β.

Nonetheless, potential similarities between TryS and protein kinases were explored. A comparison based on the SSEs surrounding the binding sites of both enzymes shows a common arrangement of helices and strands in the immediate neighbourhood. Although these similarities are probably the link between ATP grasp enzymes and kinases,[100,370,454] we showed that this SSE-based similarity is not suited to prioritise specific kinase families for a drug repurposing strategy.

Instead, the physicochemical and shape similarities between TryS and protein kinases were exploited. Based on a new consensus approach interconnecting the results of different binding site comparison methodologies, the ATP binding sites of Erk2 and PKACA were identified as the most similar to the ATP-accommodating cavities of the structural models of *Li*TryS and *Lm*TryS. These analyses could assist in the establishment of new hit compounds for the inhibition of TryS. However, the number of known inhibitors for both protein kinases in publicly available databases was low. This and further factors hampered the application of a pure drug repurposing strategy for this test case. A combined workflow using site comparisons, substructure searches, and an ensemble docking approach led to some promising hits which are worth experimental testing, given their proposed binding modes to TryS and the underlying interactions.

CONCLUSION AND OUTLOOK

Methods for *in silico* SBVS pave the way for very elaborate structure analyses and support the lead discovery and optimisation workflow.[3] Especially in the case of largely unexplored targets, such as TryS, they offer valuable means to identify interesting starting points for the development of new drugs. On the other hand, the large number of alternative methods opens up a multitude of different ways to achieve this aim. A reasonable way to find promising hits for a new target is the establishment of a drug repurposing workflow. Starting with a molecular structure of the POI, we can identify druggable ligand binding sites of the protein. Subsequently, the detected pockets can be compared to those of well-characterised therapeutically relevant proteins. Based on detected similarities, compounds which might also modulate the activity of the POI can be selected. Finally, the application of molecular docking and the subsequent visual inspection of the proposed binding modes of well-characterised drugs can help to extract the most interesting compounds. In consequence, some molecules of this selection, usually between below 1% and 25% for most VS approaches,[455] might show the desired bioactivity on the POI.

However, considering the number of approaches analysed herein, it is evident that the researcher might spend a considerably large amount of time in the literature, searching for the most appropriate tools to use. Provided that there are 60 methods for the identification of putative binding sites, 21 algorithms for the comparison of the predicted sites against known ones, at least 10 bioactivity databases,[456] and about 16 well-known docking algorithms[457] (neglecting the different options of pose scoring which are offered by these methods), there are 201,600 possibilities to pursue a drug repurposing workflow. The situation gets even worse as only subsets of the available approaches are part of this crude calculation and even new, more elaborate methods are still emerging. The question arises whether all of these conceivable routes lead to the anticipated outcome.

In consequence, it is evident that we need guidelines to enable a fast and reliable decision for the most appropriate methods. The first two topics of this thesis outline that it is possible to derive rational criteria for the choice of the most appropriate software. The results eventually show that it is necessary to provide the outcome of the methods for these criteria to the scientific community. However, they also underline the necessity of reliable and up-to-date benchmark datasets that enable an objective and target-oriented evaluation of novel methods. Only if we can ensure that both prerequisites are met, it is possible to address the challenges of an SBVS project.

With respect to the identification of binding sites, a single tool outperforms the remaining ones regarding nearly all benchmark criteria analysed herein (the hit rates, the accuracy, the robustness, the applicability to comparative models, and the prediction of cryptic and allosteric sites). Its results even outmatch these of elaborate meta approaches indicating its general suitability. So, we can highlight one superior approach which should be further optimised and evaluated in the future to ensure the reliable prediction of a broad spectrum of biologically interesting small molecule binding sites.

Things are different with respect to the comparison of binding sites. The comparison methods, which differ in the way the site is modelled, show distinct strengths and weaknesses with respect to the quality indicators analysed herein. A reliable decision for a tool depends on the focus of the study. Additionally, an even larger analysis of the multitude of available binding site comparison methods might already provide answers for the open questions of these analyses. However, the benchmark results provide the basis to derive guidelines for the reliable detection of binding site similarities.

The dependence on the desired outcome might also explain the plethora of developed methods for the assignment of SSEs in protein structures. The use of the most commonly applied methods, in this case, DSSP and STRIDE, does not always lead to the best results. This was shown for the SSE-based binding site comparison method SLOT. Kinks and irregularities in the assigned helices and strands lead to sub-optimal graph models. To meet this challenge, we developed SCOT. Retaining the robustness of hydrogen bond-based assignment methods, it assigns geometrically uniform and regular SSEs. This individual feature of the method enables a different perspective on the stable backbone conformations in proteins and their involvement in protein function. The establishment of SCOT also facilitates the automated comparison of SSE arrangements in the proximity of protein binding sites with SLOT. Its applicability as an alternative approach to conventional binding site comparison methods in the context of ligand-sensing cores could be tested for the case study on TryS.

TryS, a vital enzyme of the parasites of the order Trypanosomatida, is a therapeutically interesting target for the treatment of leishmaniasis. The known crystal structures of this ATP-grasp enzyme enabled the modelling of the substrate binding modes and the subsequent analysis of the protein dynamics. The activity of the enzyme is inhibited by a compound with scaffold-based similarities to known protein kinase inhibitors. This finding and the similarities between the ATP binding sites of TryS and kinases suggest that a drug repurposing strategy could lead to new insights into the inhibition of the POI.

Analysing the inhibitor scaffold-based relationship between TryS and protein kinases, a well-founded binding mode of paullone-derived compounds in TryS was predicted. However, using an SSE-based alignment of the ATP binding sites of the paullone-bound kinase and TryS, the binding mode differed for both enzymes. Nonetheless, the SSE-based similarities of TryS and a broad spectrum of protein kinases were analysed. A strategy to derive promising TryS inhibitors from a molecular docking of known protein kinase inhibitors failed. In consequence, it was tested whether it is possible to identify the protein kinases whose ATP binding sites are most related to those of TryS. In contrast to previous findings, the application of the ligand-sensing cores approach to the POI did not offer the possibility to reliably detect the most similar kinases. It was found that the relationship between TryS, protein kinases, and other ATP-grasp enzymes is a general one that does not allow for a prioritisation of the most similar proteins. As shown before,[458] the number of unique arrangements of SSEs (even if restricted to the ligand binding sites) is considerably small which hampers the selection of the most similar targets for a repurposing workflow.

In contrast, the analysis of the consensus of residue-, surface-, and interaction-based binding site comparison methods helped to elucidate physicochemical and shape similarities of protein kinases to TryS. The resulting finding that the ATP binding sites of two protein kinases, Erk2 and PKACA, and TryS share common residues and a comparable shape paved the way for a reliable repurposing analysis. However, the docking study with inhibitors of the protein kinases revealed also crucial dissimilarities of the ATP binding sites. Using the additional opportunity of small molecule similarity searches, it was finally possible to identify compounds which are similar to the inhibitors of both protein kinases and use them for the molecular docking into the ATP binding site of TryS.

A subset of compounds showed promising predicted binding modes in the ATP binding site of the target. A representative set of interesting small molecules with distinct interaction patterns was subsequently selected for further studies. The biochemical validation of their inhibitory potency with respect to TryS might open up new opportunities for the inhibition of this therapeutically relevant enzyme. The testing of the compounds has to be the next step to finally validate the suitability of binding site comparison in early drug discovery with respect to this case study.

Summing up all parts of this thesis, it is still a challenging task to apply SBVS approaches to largely unexplored targets. Prior to the application of a repurposing workflow to this type of proteins, an in-depth analysis of all underlying tools was inevitable. In contrast to well-characterised proteins, it is not possible to optimise the parameters of the molecular docking or to use a focused set of methods for the prediction and comparison of binding sites. Rather, it is necessary to know the best subset of generally applicable methods which can be used in this blind challenge. The success of the outcome is dependent on the performance of multiple methods which are used in combination. However, the analyses herein shed light on important criteria that have to be considered and offer guidelines for the expedient application of binding site prediction and comparison methods.

SCHLUSSFOLGERUNG UND AUSBLICK

Methoden für das strukturbasierte virtuelle Screening ebnen den Weg für umfangreiche Strukturanalysen und unterstützen die Suche nach und Optimierung von Leitstrukturen.[3] Insbesondere für kaum untersuchte Targets, wie z. B. das Enzym Trypanothionsynthetase, sind diese Methoden wertvolle Mittel zum Auffinden interessanter Ausgangspunkte für die Entwicklung neuer Medikamente. Andererseits eröffnet die große Anzahl alternativer Methoden eine Vielzahl von Möglichkeiten, um dieses Ziel zu erreichen. Eine sinnvoller Weg zu vielversprechenden Hits für ein neues Target ist die Etablierung eines Workflows zur Identifikation bekannter Wirkstoffe, die auch zur Modulation der Aktivität des zu untersuchenden Proteins anwendbar sind („drug repuposing“). Beginnend mit der molekularen Struktur des Targets können Ligandenbindetaschen identifiziert werden, die „druggable“ sind, d. h. Bindetaschen, deren Bindung an Liganden zu einer Modifikation der Funktion des Zielproteins führt. Anschließend können die detektierten Bindetaschen mit denen gut charakterisierter, therapeutisch relevanter Proteinen verglichen werden. Basierend auf den gefundenen Ähnlichkeiten können Moleküle selektiert werden, die möglicherweise auch die Aktivität des zu untersuchenden Proteins modulieren. Ein abschließendes molekulares Docking und die visuelle Analyse der vorhergesagten Bindeposen von bereits bekannten Medikamenten können helfen, die interessantesten Verbindungen zu extrahieren. Schließlich werden einige Moleküle dieser Selektion, in der Regel zwischen unter 1 % und 25 % für die meisten virtuellen Screening Ansätze,[455] die gewünschte Bioaktivität aufweisen.

Betrachtet man allerdings die Anzahl der hier untersuchten Ansätze, so ist es offensichtlich, dass ein Forscher eine beträchtliche Zeit auf der Suche nach den geeignetsten Methoden in der Literatur verbringt. Unter der Annahme, dass 60 Methoden zur Identifikation möglicher Bindetaschen zur Verfügung stehen, dass 21 Methoden den Vergleich der vorhergesagten Bindetaschen zu bekannten ermöglichen, dass es 10 Bioaktivitätsdatenbanken[456] und etwa 16 bekannte Dockingalgorithmen (unter Vernachlässigung der möglichen Optionen zur Bewertung der Bindeposen)[457] gibt, so stehen 201.600 Möglichkeiten zur Verfügung, um einen „drug repurposing“ Workflow zu etablieren. Diese Situation verschlechtert sich zusätzlich, da nur Untergruppen der verfügbaren Ansätze Teil dieser groben Berechnung sind und weiterhin neue, aufwendigere Methoden entwickelt werden. Es stellt sich daher die Frage, ob alle diese denkbaren Wege zum gewünschten Ergebnis führen.

Folglich ist es offensichtlich, dass wir Orientierungshilfen benötigen, die eine schnelle und zuverlässige Entscheidung für die passendste Methode ermöglichen. Die ersten beiden Schwerpunkte dieser Arbeit zeigen, dass eine Ableitung rationaler Kriterien zur Auswahl der geeigneten Software möglich ist. Die Ergebnisse veranschaulichen, dass es von Nutzen ist, der Wissenschaftsgemeinde die Resultate der Methoden im Bezug auf diese Qualitätskriterien zu kommunizieren. Außerdem unterstreichen die Ergebnisse die Notwendigkeit zuverlässiger und aktueller Benchmarkdatensätzen für eine objektive und zielorientierte Evaluation neuer Methoden. Nur wenn diese beiden Voraussetzungen erfüllt sind, ist es möglich, sich den Anforderungen von strukturbasieren virtuellen Screeningprojekten zu stellen.

Berücksichtigt man die Benchmarkkriterien der Analyse von Methoden zur Identifikation von Bindetaschen (Trefferraten, Genauigkeit, Robustheit, Anwendbarkeit für Homologiemodelle, und die Vorhersage kryptischer und allosterischer Bindetaschen), so ist eine Methode allen anderen evaluierten Ansätzen überlegen. Die allgemeine Anwendbarkeit dieser Software wird untermauert durch die überzeugenderen Ergebnisse im Vergleich zu kombinierten Ansätzen. Folglich ist es möglich, eine herausragende Methode zu benennen, die in der Zukunft weiter optimiert und evaluiert werden sollte, um eine zuverlässige Vorhersage eines breiten Spektrums biologisch interessanter Bindetaschen kleiner Moleküle zu garantieren.

Im Bezug auf den Vergleich von Proteinbindetaschen waren die Analyseergebnisse weniger eindeutig. Die Vergleichsmethoden, die sich in der Modellierung der Bindetasche unterscheiden, weisen im Hinblick auf die hier untersuchten Qualitätsindikatoren spezifische Stärken und Schwächen auf. Eine zuverlässige Wahl eines Tools hängt vom Schwerpunkt einer Studie ab. Zusätzlich könnte die Analyse einer noch größeren Auswahl von Ansätzen bereits einige offene, sich aus den Analysen ergebende Fragen beantworten. Die ersten Benchmarkergebnisse liefern jedoch die Grundlage zur Ableitung von Leitlinien für die zuverlässige Identifikation von Bindetaschenähnlichkeiten.

Die Abhängigkeit vom erwünschten Resultat könnte ebenfalls die große Menge von entwickelten Methoden für die Zuordnung von Sekundärstrukturelementen in Proteinstrukturen erklären. Die Verwendung der gängigsten Methoden, in diesem Fall DSSP und STRIDE, führt nicht immer zu den besten Ergebnissen. Das konnte für die Sekundärstrukturelement-basierte Software SLOT zum Vergleich von Bindetaschen gezeigt werden. Knicke und Unregelmäßigkeiten in den zugeordneten Helices und Strands führen zu suboptimalen Graphmodellen. Um dieser Herausforderung gerecht zu werden, haben wir SCOT entwickelt. Die Software ordnet geometrisch regelmäßige Sekundärstrukturelemente zu, ist aber ebenso robust wie andere Wasserstoffbrücken-basierte Methoden zur Zuweisung von Sekundärstrukturelementen. Diese individuelle Eigenschaft der Methode ermöglicht einen anderen Blickwinkel auf stabile Proteinrückgratkonformationen und deren Bedeutung für die Funktion eines Proteins. Die Entwicklung von SCOT erleichtert zusätzlich den automatisierten Vergleich von Sekundärstrukturanordnungen in Proteinbindetaschen mit SLOT. Dessen Anwendbarkeit hinsichtlich gemeinsamer „ligand-sensing cores“ als Alternative zu konventionellen Methoden zum Bindetaschenvergleich konnte anhand der Fallstudie an Trypanothionsynthetase getestet werden.

Trypanothionsynthetase, ein essenzielles Enzym für Parasiten der Ordnung Trypanosomatida, ist ein therapeutisch interessantes Target zur Behandlung von Leishmaniose. Die bekannten Kristallstrukturen dieses ATP-Grasp Enzyms ermöglichten die Modellierung der Bindemodi der Substrate und die nachfolgende Analyse der Proteindynamik. Die Aktivität des Enzyms kann durch Verbindungen mit Grundgerüst-basierten Ähnlichkeiten zu bekannten Proteinkinase-inhibitoren gehemmt werden. Diese Erkenntnis und die Ähnlichkeiten zwischen den ATP-Bindestellen von Trypanothionsynthetase und Proteinkinasen deuten drauf hin, dass eine “drug repurposing” Strategie zu neuen Einblicken in die Inhibition des Targets führen könnte.

Bezüglich der Analyse der Beziehung zwischen Trypanothionsynthetase und Proteinkinasen konnte ein gut begründeter Bindemodus von Verbindungen, die vom gemeinsamen Inhibitorgrundgerüst der Paullone abgeleitet sind, vorhergesagt werden. Legt man allerdings die Sekundärstrukturelement-basierte Überlagerung beider Enzyme zugrunde, so unterscheiden sich die Bindemodi der Paullone in Trypanothionsynthetase und Kinasen. Nichtsdestotrotz wurden die Sekundärstruktur-basierten Ähnlichkeiten zwischen dem Target und einem breiten Spektrum von Proteinkinasen weiterführend untersucht.

Eine Strategie zur Identifikation vielversprechender Trypanothionsynthetaseinhibitoren basierend auf einem molekularen Docking bekannter Kinaseinhbitoren schlug fehl. In der Folge wurde evaluiert, ob Proteinkinasen identifiziert werden können, deren ATP-Bindetaschen am ähnlichsten zur entsprechenden Bindetasche in Trypanothionsythetase sind. Im Gegensatz zu vorherigen Erkenntnissen war die zuverlässig Anwendung des "ligand-sensing cores" Ansatzes zur Identifikation der größten Ähnlichkeiten zwischen Proteinkinasen und dem zu untersuchenden Protein nicht möglich. Die Sekundärstrukturelement-basierte Beziehung zwischen Trypanothionsynthetase, Proteinkinasen und anderen ATP-Grasp Enzymen ist von allgemeiner Natur und erlaubt keine Priorisierung der ähnlichsten Proteine. Wie bereits gezeigt,[458] ist die Anzahl einzigartiger Anordnungen von Sekundärstruktur-elementen (selbst unter Beschränkung auf die Bindetaschen) äußerst gering und verhindert die Selektion der ähnlichsten Targets im Rahmen eines „drug repurposing" Workflows.

Im Gegensatz dazu half die Analyse des Konsensus von Reste-, Oberflächen- und Interaktions-basierten Bindetaschenvergleichsmethoden bei der Aufklärung von Ähnlichkeiten zwischen Proteinkinasen und Trypanothionsythetase im Hinblick auf die physikochemischen Eigenschaften und die Form der Bindetaschen. Die resultierende Erkenntnis, dass die ATP-Bindetaschen des Targets und der beiden Proteinkinasen Erk2 und PKACA gemeinsame Reste und eine gemeinsame Form aufweisen, ebnete einer „drug repurposing" Analyse den Weg. Allerdings zeigte das initiale molekulare Docking bekannter Inhibitoren beider Kinasen entscheidende Unterschiede der ATP-Bindetaschen auf. Unter Nutzung der zusätzlichen Möglichkeit einer Ähnlichkeitssuche war es schließlich gelungen, Verbindungen zu identifizieren, die ähnlich zu den Inhibitoren beider Proteinkinasen sind und diese für ein molekulares Docking in die ATP-Bindestelle von Trypanothionsynthetase zu nutzen.

Eine Untergruppe von Verbindungen wies vielversprechende vorhergesagte Bindemodi in der ATP-Bindetasche des Targets auf. Eine repräsentative Auswahl von interessanten Molekülen mit eindeutigen gemeinsamen Interaktionsmustern wurde nachfolgend für weitere Studien selektiert. Die biochemische Validierung des inhibitorischen Potenzials gegenüber des Enzyms Trypanothionsynthetase könnte neue Möglichkeiten zur Inhibition dieses therapeutisch relevanten Targets eröffnen. Die Testung der Verbindungen muss der nächste Schritt sein, um abschließend die Eignung von Bindetaschenvergleichen als möglichen Startpunkt in der frühen Wirkstoffentwicklung im Hinblick auf diese Fallstudie zu validieren.

Unter Berücksichtigung aller Teile dieser Arbeit lässt sich zusammenfassend feststellen, dass es nach wie vor eine anspruchsvolle Aufgabe ist, Ansätze des strukturbasierten virtuellen Screenings auf größtenteils unerforschte Targets anzuwenden. Vor der eigentlichen Anwendung einer „drug repurposing“ Strategie für diese Art von Proteinen ist eine intensive Analyse der zugrunde liegenden Methoden unerlässlich. Im Vergleich zu gut charakterisierten Proteinen ist es nicht möglich, die Parameter des molekularen Dockings zu optimieren oder eine zielorientierte Zusammenstellung von Methoden für die Vorhersage und den Vergleich von Proteinbindetaschen zu verwenden. Vielmehr ist es nötig, die beste Teilmenge der allgemein anwendbaren Software zu kennen und in dieser „Blindstudie“ zu verwenden. Der Erfolg des Ergebnisses hängt von der Qualität zahlreicher Methoden ab, die in Kombination verwendet werden. Jedoch zeigen die Analysen dieser Arbeit wichtige zu berücksichtigende Kriterien auf und bieten Orientierungshilfen für eine zweckdienliche Auswahl und Anwendung von Methoden für die Vorhersage und den Vergleich von Bindetaschen.

REFERENCES

[1] World Health Organization. Neglected tropical diseases. http://www.who.int/neglected_diseases/diseases/en/ (accessed November 3, 2018).

[2] Barrett, M. P.; Croft, S. L. Management of trypanosomiasis and leishmaniasis. *Br. Med. Bull.* **2012**, *104*, 175–196.

[3] Lionta, E.; Spyrou, G.; Vassilatis, D. K.; Cournia, Z. Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Curr. Top. Med. Chem.* **2014**, *14* (16), 1923–1938.

[4] Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10* (9), 787–797.

[5] Sippl, W.; Höltje, H.-D. Structure-based 3D-QSAR - Merging the accuracy of structure-based alignments with the computational efficiency of ligand-based methods. *J. Mol. Struct.: THEOCHEM* **2000**, *503* (1-2), 31–50.

[6] Petrey, D.; Honig, B. Protein structure prediction: Inroads to biology. *Mol. Cell* **2005**, *20* (6), 811–819.

[7] Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

[8] Schmidtke, P.; Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **2010**, *53* (15), 5858–5867.

[9] Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1* (9), 727–730.

[10] Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: An annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46* (2), 717–727.

[11] Brady, G. P.; Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14* (4), 383–401.

[12] Huang, B.; Schroeder, M. LIGSITE^{CSC}: Predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19.

[13] Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849–857.

[14] Zhong, S.; MacKerell, A. D. Binding Response: A descriptor for selecting ligand binding site on protein surfaces. *J. Chem. Inf. Model.* **2007**, *47* (6), 2303–2315.

[15] Gao, M.; Skolnick, J. A comprehensive survey of small-molecule binding pockets in proteins. *PLoS Comput. Biol.* **2013**, *9* (10), e1003302.

[16] Glaser, F.; Pupko, T.; Paz, I.; Bell, R. E.; Bechor-Shental, D.; Martz, E.; Ben-Tal, N. ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **2003**, *19* (1), 163–164.

[17] Bianchi, V.; Gherardini, P. F.; Helmer-Citterich, M.; Ausiello, G. Identification of binding pockets in protein structures using a knowledge-based potential derived from local structural similarities. *BMC Bioinf.* **2012**, *13 Suppl 4*, S17.

[18] Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; Fraser, J. S.; Sali, A. CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J. Mol. Biol.* **2016**, *428* (4), 709–719.

[19] Kim, J.-K.; Cho, Y.; Laskowski, R. A.; Ryu, S. E.; Sugihara, K.; Kim, D.-S. BetaVoid: Molecular voids via β-complexes and Voronoi diagrams. *Proteins* **2014**, *82* (9), 1829–1849.

[20] Kleywegt, G. J.; Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *50* (Pt 2), 178–185.

[21] Neuvirth, H.; Raz, R.; Schreiber, G. ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **2004**, *338* (1), 181–199.

[22] Guo, F.; Li, S. C.; Wang, L.; Zhu, D. Protein-protein binding site identification by enumerating the configurations. *BMC Bioinf.* **2012**, *13*, 158.

[23] Paz, I.; Kligun, E.; Bengad, B.; Mandel-Gutfreund, Y. BindUP: A web server for non-homology-based prediction of DNA and RNA binding proteins. *Nucleic Acids Res.* **2016**, *44* (W1), W568-74.

[24] Ahmad, S.; Gromiha, M. M.; Sarai, A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* **2004**, *20* (4), 477–486.

[25] Parca, L.; Ferré, F.; Ausiello, G.; Helmer-Citterich, M. Nucleos: A web server for the identification of nucleotide-binding sites in protein structures. *Nucleic Acids Res.* **2013**, *41* (Web Server issue), W281-5.

[26] Sun, J.; Chen, K. NSiteMatch: Prediction of binding sites of nucleotides by identifying the structure similarity of local surface patches. *Comput. Math. Methods Med.* **2017**, *2017*, 5471607.

[27] Hu, X.; Dong, Q.; Yang, J.; Zhang, Y. Recognizing metal and acid radical ion-binding sites by integrating *ab initio* modeling with template-based transferals. *Bioinformatics* **2016**, *32* (21), 3260–3269.

[28] Lippi, M.; Passerini, A.; Punta, M.; Rost, B.; Frasconi, P. MetalDetector: A web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics* **2008**, *24* (18), 2094–2095.

[29] Babor, M.; Gerzon, S.; Raveh, B.; Sobolev, V.; Edelman, M. Prediction of transition metal-binding sites from apo protein structures. *Proteins* **2008**, *70* (1), 208–217.

[30] Parca, L.; Mangone, I.; Gherardini, P. F.; Ausiello, G.; Helmer-Citterich, M. Phosfinder: A web server for the identification of phosphate-binding sites on protein structures. *Nucleic Acids Res.* **2011**, *39* (Web Server issue), W278-82.

[31] Goncearenco, A.; Mitternacht, S.; Yong, T.; Eisenhaber, B.; Eisenhaber, F.; Berezovsky, I. N. SPACER: Server for predicting allosteric communication and effects of regulation. *Nucleic Acids Res.* **2013**, *41* (Web Server issue), W266-72.

[32] Greener, J. G.; Sternberg, M. J. E. AlloPred: Prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinf.* **2015**, *16*, 335.

[33] Song, K.; Liu, X.; Huang, W.; Lu, S.; Shen, Q.; Zhang, L.; Zhang, J. Improved method for the identification and validation of allosteric sites. *J. Chem. Inf. Model.* **2017**, *57* (9), 2358–2363.

[34] Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.* **2009**, *5* (12), e1000585.

[35] Huang, B. MetaPocket: A meta approach to improve protein ligand binding site prediction. *OMICS* **2009**, *13* (4), 325–330.

[36] Yang, J.; Roy, A.; Zhang, Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29* (20), 2588–2595.

[37] Broomhead, N. K.; Soliman, M. E. Can we rely on computational predictions to correctly identify ligand binding sites on novel protein drug targets? Assessment of binding site prediction methods and a protocol for validation of predicted binding sites. *Cell Biochem. Biophys.* **2017**, *75* (1), 15–23.

[38] Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein clefts in molecular recognition and function. *Protein Sci.* **1996**, *5* (12), 2438–2452.

[39] Gao, J.; Liu, Q.; Kang, H.; Cao, Z.; Zhu, R. Comparison of different ranking methods in protein-ligand binding site prediction. *Int. J. Mol. Sci.* **2012**, *13* (7), 8752–8761.

[40] Wong, G. Y.; Leung, F. H. F.; Ling, S. H. Predicting protein-ligand binding site using support vector machine with protein properties. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2013**, *10* (6), 1517–1529.

[41] Krivák, R.; Hoksza, D. Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. *J. Cheminf.* **2015**, *7*, 12.

[42] Schmidtke, P.; Souaille, C.; Estienne, F.; Baurin, N.; Kroemer, R. T. Large-scale comparison of four binding site detection algorithms. *J. Chem. Inf. Model.* **2010**, *50* (12), 2191–2200.

[43] Chen, K.; Mizianty, M. J.; Gao, J.; Kurgan, L. A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure* **2011**, *19* (5), 613–621.

[44] Leis, S.; Schneider, S.; Zacharias, M. *In silico* prediction of binding sites on proteins. *Curr. Med. Chem.* **2010**, *17* (15), 1550–1562.

[45] Dukka, B. K. Structure-based methods for computational protein functional site prediction. *Comput. Struct. Biotechnol. J.* **2013**, *8*, e201308005.

[46] Xie, Z.-R.; Hwang, M.-J. Methods for predicting protein-ligand binding sites. *Methods Mol. Biol.* **2015**, *1215*, 383–398.

[47] Krone, M.; Kozlíková, B.; Lindow, N.; Baaden, M.; Baum, D.; Parulek, J.; Hege, H.-C.; Viola, I. Visual analysis of biomolecular cavities: State of the art. *Comput. Graph. Forum* **2016**, *35* (3), 527–551.

[48] Ehrt, C.; Brinkjost, T.; Koch, O. Impact of binding site comparisons on medicinal chemistry and rational molecular design. *J. Med. Chem.* **2016**, *59* (9), 4121–4151.

[49] Yeturu, K.; Chandra, N. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, 543.

[50] Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **2012**, *52* (8), 2287–2299.
[51] Volkamer, A.; Rarey, M. Exploiting structural information for drug-target assessment. *Future Med. Chem.* **2014**, *6* (3), 319–331.
[52] Konc, J.; Janežič, D. Binding site comparison for function prediction and pharmaceutical discovery. *Curr. Opin. Struct. Biol.* **2014**, *25*, 34–39.
[53] Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4* (3), 209–220.
[54] Haupt, V. J.; Daminelli, S.; Schroeder, M. Drug promiscuity in PDB: Protein binding site similarity is key. *PLoS ONE* **2013**, *8* (6), e65894.
[55] Chartier, M.; Morency, L.-P.; Zylber, M. I.; Najmanovich, R. J. Large-scale detection of drug off-targets: Hypotheses for drug repurposing and understanding side-effects. *BMC Pharmacol. Toxicol.* **2017**, *18* (1), 18.
[56] Wu, C. Y.; Chen, Y. C.; Lim, C. A structural-alphabet-based strategy for finding structural motifs across protein families. *Nucleic Acids Res.* **2010**, *38* (14), e150.
[57] Najmanovich, R. J.; Kurbatova, N.; Thornton, J. M. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics* **2008**, *24* (16), i105-i111.
[58] Roy, A.; Yang, J.; Zhang, Y. COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **2012**, *40* (Web Server issue), 7.
[59] Kinoshita, K.; Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **2003**, *12* (8), 1589–1595.
[60] Reisen, F.; Weisel, M.; Kriegl, J. M.; Schneider, G. Self-organizing fuzzy graphs for structure-based comparison of protein pockets. *J. Proteome Res.* **2010**, *9* (12), 6498–6510.
[61] Konc, J.; Janežič, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26* (9), 1160–1168.
[62] Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339* (3), 607–633.
[63] Binkowski, T. A.; Joachimiak, A. Protein functional surfaces: Global shape matching and local spatial alignments of ligand binding sites. *BMC Struct. Biol.* **2008**, *8*, 45.
[64] Schmitt, S.; Hendlich, M.; Klebe, G. From structure to function: A new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angew. Chem. Int. Ed.* **2001**, *40* (17), 3141–3144.
[65] Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323* (2), 387–406.
[66] Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47* (1), 195–207.
[67] Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P. Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **2006**, *65* (1), 124–135.

[68] Koch, M. A.; Wittenberg, L.-O.; Basu, S.; Jeyaraj, D. A.; Gourzoulidou, E.; Reinecke, K.; Odermatt, A.; Waldmann, H. Compound library development guided by protein structure similarity clustering and natural product structure. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (48), 16721–16726.

[69] Xie, L.; Xie, L.; Bourne, P. E. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics* **2009**, *25* (12), 12.

[70] Xie, L.; Bourne, P. E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (14), 5441–5446.

[71] Zhang, Y.; Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33* (7), 2302–2309.

[72] Totrov, M. Ligand binding site superposition and comparison based on Atomic Property Fields: Identification of distant homologues, convergent evolution and PDB-wide clustering of binding sites. *BMC Bioinf.* **2011**, *12 Suppl 1*, S35.

[73] Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50* (1), 123–135.

[74] Wood, D. J.; Vlieg, J. de; Wagener, M.; Ritschel, T. Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. *J. Chem. Inf. Model.* **2012**, *52* (8), 2031–2043.

[75] Brakoulias, A.; Jackson, R. M. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins* **2004**, *56* (2), 250–260.

[76] Spitzer, R.; Cleves, A. E.; Jain, A. N. Surface-based protein binding pocket similarity. *Proteins* **2011**, *79* (9), 2746–2763.

[77] Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71* (4), 1755–1778.

[78] Liu, T.; Altman, R. B. Using multiple microenvironments to find similar ligand-binding sites: Application to kinase inhibitor binding. *PLoS Comput. Biol.* **2011**, *7* (12), e1002326.

[79] Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47* (2), 279–294.

[80] Das, S.; Kokardekar, A.; Breneman, C. M. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model.* **2009**, *49* (12), 2863–2872.

[81] Li, G.-H.; Huang, J.-F. CMASA: An accurate algorithm for detecting local protein structural similarity and its application to enzyme catalytic site annotation. *BMC Bioinf.* **2010**, *11*, 439.

[82] Pauling, L.; Corey, R. B.; Branson, H. R. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **1951**, *37* (4), 205–211.

[83] Nagy, G.; Oostenbrink, C. Dihedral-based segment identification and classification of biopolymers I: Proteins. *J. Chem. Inf. Model.* **2014**, *54* (1), 266–277.

[84] Donohue, J. Hydrogen bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **1953**, *39* (6), 470–478.
[85] Bamford, C. H.; Elliot, A.; Hanby, W. E. *Synthetic Polypeptides: Preparation, structure, and properties*; Academic Press: New York, NY, 1956.
[86] Low, B. W.; Grenville-Wells, H. J. Generalized mathematical relationships for polypeptide chain helices: The coordinates of the II helix. *Proc. Natl. Acad. Sci. U. S. A.* **1953**, *39* (8), 785–801.
[87] Arnott, S.; Dover, S.D.; Elliott, A. Structure of β-poly-l-alanine: Refined atomic coordinates for an anti-parallel β-pleated sheet. *J. Mol. Biol.* **1967**, *30* (1), 201–208.
[88] Schellman, J. A.; Schellman, C. The Conformation of Polypeptide Chains in Proteins. *The Proteins Composition, Structure, and Function*; Elsevier, 1964; pp 1–137.
[89] Adzhubei, A. A.; Eisenmenger, F.; Tumanyan, V. G.; Zinke, M.; Brodzinski, S.; Esipova, N. G. Approaching a complete classification of protein secondary structure. *J. Biomol. Struct. Dyn.* **1987**, *5* (3), 689–704.
[90] Novotny, M.; Kleywegt, G. J. A survey of left-handed helices in protein structures. *J. Mol. Biol.* **2005**, *347* (2), 231–241.
[91] Hollingsworth, S. A.; Berkholz, D. S.; Karplus, P. A. On the occurrence of linear groups in proteins. *Protein Sci.* **2009**, *18* (6), 1321–1325.
[92] Narwani, T. J.; Santuz, H.; Shinada, N.; Melarkode Vattekatte, A.; Ghouzam, Y.; Srinivasan, N.; Gelly, J.-C.; de Brevern, A. G. Recent advances on polyproline II. *Amino Acids* **2017**, *49* (4), 705–713.
[93] Koch, O. Use of secondary structure element information in drug design: Polypharmacology and conserved motifs in protein-ligand binding and protein-protein interfaces. *Future Med. Chem.* **2011**, *3* (6), 699–708.
[94] Forsyth, W. R.; Antosiewicz, J. M.; Robertson, A. D. Empirical relationships between protein structure and carboxyl pKa values in proteins. *Proteins* **2002**, *48* (2), 388–403.
[95] Miranda, J. J. L. Position-dependent interactions between cysteine residues and the helix dipole. *Protein Sci.* **2003**, *12* (1), 73–81.
[96] Kumar, P.; Bansal, M. Dissecting π-helices: Sequence, structure and function. *FEBS J.* **2015**, *282* (22), 4415–4432.
[97] Cooley, R. B.; Arp, D. J.; Karplus, P. A. Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices that enhance protein functionality. *J. Mol. Biol.* **2010**, *404* (2), 232–246.
[98] Enkhbayar, P.; Hikichi, K.; Osaki, M.; Kretsinger, R. H.; Matsushima, N. 3_{10}-helices in proteins are parahelices. *Proteins* **2006**, *64* (3), 691–699.
[99] Vieira-Pires, R. S.; Morais-Cabral, J. H. 3_{10} helices in channels and other membrane proteins. *J. Gen. Physiol.* **2010**, *136* (6), 585–592.
[100] Grishin, N. V. Phosphatidylinositol phosphate kinase: A link between protein kinase and glutathione synthase folds. *J. Mol. Biol.* **1999**, *291* (2), 239–247.
[101] Stepanenko, O. V.; Kuznetsova, I. M.; Verkhusha, V. V.; Turoverov, K. K. β-barrel scaffold of fluorescent proteins: Folding, stability and role in chromophore formation. *Int. Rev. Cell Mol. Biol.* **2013**, *302*, 221–278.

[102] Riccardi, L.; Mereghetti, P. Induced fit in protein multimerization: The HFBI case. *PLoS Comput. Biol.* **2016**, *12* (11), e1005202.
[103] Norrby, E. Prions and protein-folding diseases. *J. Intern. Med.* **2011**, *270* (1), 1–14.
[104] Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22* (12), 2577–2637.
[105] Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **1995**, *23* (4), 566–579.
[106] Koch, O.; Cole, J. C. An automated method for consistent helix assignment using turn information. *Proteins* **2011**, *79* (5), 1416–1426.
[107] Fodje, M. N.; Al-Karadaghi, S. Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Eng., Des. Sel.* **2002**, *15* (5), 353–358.
[108] Hutchinson, E. G.; Thornton, J. M. PROMOTIF - A program to identify and analyze structural motifs in proteins. *Protein Sci.* **1996**, *5* (2), 212–220.
[109] Overington, J. P.; Johnson, M. S.; Sali, A.; Blundell, T. L. Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. *Proc. R. Soc. B* **1990**, *241* (1301), 132–145.
[110] Mizuguchi, K.; Deane, C. M.; Blundell, T. L.; Johnson, M. S.; Overington, J. P. JOY: Protein sequence-structure representation and analysis. *Bioinformatics* **1998**, *14* (7), 617–623.
[111] Andersen, C. A. F.; Palmer, A. G.; Brunak, S.; Rost, B. Continuum secondary structure captures protein flexibility. *Structure* **2002**, *10* (2), 175–184.
[112] Zacharias, J.; Knapp, E.-W. Protein secondary structure classification revisited: Processing DSSP information with PSSC. *J. Chem. Inf. Model.* **2014**, *54* (7), 2166–2179.
[113] Mansiaux, Y.; Joseph, A. P.; Gelly, J.-C.; de Brevern, A. G. Assignment of polyproline II conformation and analysis of sequence-structure relationship. *PLoS ONE* **2011**, *6* (3), e18401.
[114] Chen, K.-H.; Huang, G.-S.; Lu, C. L.; Lee, R. C.-T. A multidimensional divide-and-conquer algorithm for assigning secondary structures in proteins. *Proceedings of the 26th Workshop on Combinatorial Mathematics and Computation Theory* **2009**, 1–8.
[115] Levitt, M.; Greer, J. Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.* **1977**, *114* (2), 181–239.
[116] Parisien, M.; Major, F. A new catalog of protein β-sheets. *Proteins* **2005**, *61* (3), 545–558.
[117] Srinivasan, R.; Rose, G. D. A physical basis for protein secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (25), 14258–14263.
[118] Martin, J.; Letellier, G.; Marin, A.; Taly, J.-F.; de Brevern, A. G.; Gibrat, J.-F. Protein secondary structure assignment revisited: A detailed analysis of different assignment methods. *BMC Struct. Biol.* **2005**, *5*, 17.
[119] Cubellis, M. V.; Cailliez, F.; Lovell, S. C. Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinf.* **2005**, *6 Suppl 4*, S8.
[120] Salawu, E. O. RaFoSA: Random forests secondary structure assignment for coarse-grained and all-atom protein systems. *Cogent Biol.* **2016**, *2* (1), 2010.
[121] Cao, C.; Wang, G.; Liu, A.; Xu, S.; Wang, L.; Zou, S. A new secondary structure assignment algorithm using Cα backbone fragments. *Int. J. Mol. Sci.* **2016**, *17* (3), 333.

[122] Kumar, P.; Bansal, M. Identification of local variations within secondary structures of proteins. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2015**, *71* (Pt 5), 1077–1086.
[123] Kneller, G. R.; Hinsen, K. Protein secondary-structure description with a coarse-grained model. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2015**, *71* (Pt 7), 1411–1422.
[124] Law, S. M.; Frank, A. T.; Brooks, C. L. PCASSO: A fast and efficient Cα-based method for accurately assigning protein secondary structure elements. *J. Comput. Chem.* **2014**, *35* (24), 1757–1761.
[125] Konagurthu, A. S.; Lesk, A. M.; Allison, L. Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics* **2012**, *28* (12), i97-105.
[126] Park, S. Y.; Yoo, M.-J.; Shin, J.; Cho, K.-H. SABA (Secondary structure Assignment program Based on only Alpha carbons): A novel pseudo center geometrical criterion for accurate assignment of protein secondary structures. *BMB Rep.* **2011**, *44* (2), 118–122.
[127] Konagurthu, A. S.; Allison, L.; Stuckey, P. J.; Lesk, A. M. Piecewise linear approximation of protein structures using the principle of minimum message length. *Bioinformatics* **2011**, *27* (13), i43-51.
[128] Hosseini, S.-R.; Sadeghi, M.; Pezeshk, H.; Eslahchi, C.; Habibi, M. PROSIGN: A method for protein secondary structure assignment based on three-dimensional coordinates of consecutive Cα atoms. *Comput. Biol. Chem.* **2008**, *32* (6), 406–411.
[129] Majumdar, I.; Krishna, S. S.; Grishin, N. V. PALSSE: A program to delineate linear secondary structural elements from protein structures. *BMC Bioinf.* **2005**, *6*, 202.
[130] Taylor, T.; Rivera, M.; Wilson, G.; Vaisman, I. I. New method for protein secondary structure assignment based on a simple topological descriptor. *Proteins* **2005**, *60* (3), 513–524.
[131] Dupuis, F.; Sadoc, J.-F.; Mornon, J.-P. Protein secondary structure assignment through Voronoï tessellation. *Proteins* **2004**, *55* (3), 519–528.
[132] Taylor, W. R. Defining linear segments in protein structure. *J. Mol. Biol.* **2001**, *310* (5), 1135–1150.
[133] King, S. M.; Johnson, W. C. Assigning secondary structure from protein coordinate data. *Proteins* **1999**, *35* (3), 313–320.
[134] Labesse, G.; Colloc'h, N.; Pothier, J.; Mornon, J.-P. P-SEA: A new efficient assignment of secondary structure from Cα trace of proteins. *CABIOS, Comput. Appl. Biosci.* **1997**, *13* (3), 291–295.
[135] Kleywegt, G. J.; Jones, T. A. Detecting folding motifs and similarities in protein structures. *Methods Enzymol.* **1997**, *277*, 525–545.
[136] Sklenar, H.; Etchebest, C.; Lavery, R. Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins* **1989**, *6* (1), 46–60.
[137] Richards, F. M.; Kundrot, C. E. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *Proteins* **1988**, *3* (2), 71–84.
[138] Zhang, W.; Dunker, A. K.; Zhou, Y. Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins* **2008**, *71* (1), 61–67.

[139] Colloc'h, N.; Etchebest, C.; Thoreau, E.; Henrissat, B.; Mornon, J.-P. Comparison of three algorithms for the assignment of secondary structure in proteins: The advantages of a consensus assignment. *Protein Eng.* **1993**, *6* (4), 377–382.

[140] Ma, J.; Wang, S. Algorithms, applications, and challenges of protein structure alignment. *Adv. Protein Chem. Struct. Biol.* **2014**, *94*, 121–175.

[141] Hanukoglu, I. Proteopedia: Rossmann fold: A β-α-β fold at dinucleotide binding sites. *Biochem. Mol. Biol. Educ.* **2015**, *43* (3), 206–209.

[142] Sillitoe, I.; Lewis, T. E.; Cuff, A.; Das, S.; Ashford, P.; Dawson, N. L.; Furnham, N.; Laskowski, R. A.; Lee, D.; Lees, J. G.; Lehtinen, S.; Studer, R. A.; Thornton, J. M.; Orengo, C. A. CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* **2015**, *43* (Database issue), 81.

[143] Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Mistry, J.; Mitchell, A. L.; Potter, S. C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; Salazar, G. A.; Tate, J.; Bateman, A. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44* (D1), D279-85.

[144] Fox, N. K.; Brenner, S. E.; Chandonia, J.-M. SCOPe: Structural Classification Of Proteins - extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **2014**, *42* (Database issue), 9.

[145] Cheng, H.; Schaeffer, R. D.; Liao, Y.; Kinch, L. N.; Pei, J.; Shi, S.; Kim, B.-H.; Grishin, N. V. ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.* **2014**, *10* (12), e1003926.

[146] Balamurugan, R.; Dekker, F. J.; Waldmann, H. Design of compound libraries based on natural product scaffolds and protein structure similarity clustering (PSSC). *Mol. BioSyst.* **2005**, *1* (1), 36–45.

[147] Holm, L.; Rosenström, P. Dali server: Conservation mapping in 3D. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W545-9.

[148] Dekker, F. J.; Rocks, O.; Vartak, N.; Menninger, S.; Hedberg, C.; Balamurugan, R.; Wetzel, S.; Renner, S.; Gerauer, M.; Schölermann, B.; Rusch, M.; Kramer, J. W.; Rauh, D.; Coates, G. W.; Brunsveld, L.; Bastiaens, P. I. H.; Waldmann, H. Small-molecule inhibition of APT1 affects Ras localization and signaling. *Nat. Chem. Biol.* **2010**, *6* (6), 449–456.

[149] Willmann, D.; Lim, S.; Wetzel, S.; Metzger, E.; Jandausch, A.; Wilk, W.; Jung, M.; Forne, I.; Imhof, A.; Janzer, A.; Kirfel, J.; Waldmann, H.; Schule, R.; Buettner, R. Impairment of prostate cancer cell growth by a selective and reversible lysine-specific demethylase 1 inhibitor. *Int. J. Cancer* **2012**, *131* (11), 2704–2709.

[150] Walker, D. M.; Oghumu, S.; Gupta, G.; McGwire, B. S.; Drew, M. E.; Satoskar, A. R. Mechanisms of cellular invasion by intracellular parasites. *Cell. Mol. Life Sci.* **2014**, *71* (7), 1245–1263.

[151] Steverding, D. The development of drugs for treatment of sleeping sickness: A historical review. *Parasites Vectors* **2010**, *3* (1), 15.

[152] Wilkinson, S. R.; Kelly, J. M. Trypanocidal drugs: Mechanisms, resistance and new targets. *Expert Rev. Mol. Med.* **2009**, *11*, e31.

[153] Zijlstra, E. E.; Musa, A. M.; Khalil, E. A. G.; el-Hassan, I. M.; el-Hassan, A. M. Post-kala-azar dermal leishmaniasis. *Lancet Infect. Dis.* **2003**, *3* (2), 87–98.

[154] Akhoundi, M.; Kuhls, K.; Cannet, A.; Votýpka, J.; Marty, P.; Delaunay, P.; Sereno, D. A historical overview of the classification, evolution, and dispersion of Leishmania parasites and sandflies. *PLoS Neglected Trop. Dis.* **2016**, *10* (3), e0004349.

[155] Wallace, F. G. The trypanosomatid parasites of insects and arachnids. *Exp. Parasitol.* **1966**, *18* (1), 124–193.

[156] Irigoín, F.; Cibils, L.; Comini, M. A.; Wilkinson, S. R.; Flohé, L.; Radi, R. Insights into the redox biology of *Trypanosoma cruzi*: Trypanothione metabolism and oxidant detoxification. *Free Radical Biol. Med.* **2008**, *45* (6), 733–742.

[157] Sousa, A. F.; Gomes-Alves, A. G.; Benítez, D.; Comini, M. A.; Flohé, L.; Jäger, T.; Passos, J.; Stuhlmann, F.; Tomás, A. M.; Castro, H. Genetic and chemical analyses reveal that trypanothione synthetase but not glutathionylspermidine synthetase is essential for *Leishmania infantum*. *Free Radical Biol. Med.* **2014**, *73*, 229–238.

[158] Wyllie, S.; Oza, S. L.; Patterson, S.; Spinks, D.; Thompson, S.; Fairlamb, A. H. Dissecting the essentiality of the bifunctional trypanothione synthetase-amidase in *Trypanosoma brucei* using chemical and genetic methods. *Mol. Microbiol.* **2009**, *74* (3), 529–540.

[159] Henderson, G. B.; Yamaguchi, M.; Novoa, L.; Fairlamb, A. H.; Cerami, A. Biosynthesis of the trypanosomatid metabolite trypanothione: Purification and characterization of trypanothione synthetase from *Crithidia fasciculata*. *Biochemistry* **1990**, *29* (16), 3924–3929.

[160] Fyfe, P. K.; Oza, S. L.; Fairlamb, A. H.; Hunter, W. N. Leishmania trypanothione synthetase-amidase structure reveals a basis for regulation of conflicting synthetic and hydrolytic activities. *J. Biol. Chem.* **2008**, *283* (25), 17672–17680.

[161] Sigrist, C. J. A.; Castro, E. de; Cerutti, L.; Cuche, B. A.; Hulo, N.; Bridge, A.; Bougueleret, L.; Xenarios, I. New and continuing developments at PROSITE. *Nucleic Acids Res.* **2013**, *41* (Database issue), D344-7.

[162] Koch, O.; Cappel, D.; Nocker, M.; Jäger, T.; Flohé, L.; Sotriffer, C. A.; Selzer, P. M. Molecular dynamics reveal binding mode of glutathionylspermidine by trypanothione synthetase. *PLoS ONE* **2013**, *8* (2), e56788.

[163] Oza, S. L.; Shaw, M. P.; Wyllie, S.; Fairlamb, A. H. Trypanothione biosynthesis in *Leishmania major*. *Mol. Biochem. Parasitol.* **2005**, *139* (1), 107–116.

[164] Oza, S. L. Properties of trypanothione synthetase from *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **2003**, *131* (1), 25–33.

[165] Oza, S. L.; Tetaud, E.; Ariyanayagam, M. R.; Warnon, S. S.; Fairlamb, A. H. A single enzyme catalyses formation of trypanothione from glutathione and spermidine in *Trypanosoma cruzi*. *J. Biol. Chem.* **2002**, *277* (39), 35853–35861.

[166] Bollinger, J. M.; Kwon, D. S.; Huisman, G. W.; Kolter, R.; Walsh, C. T. Glutathionylspermidine metabolism in *Escherichia coli*. *J. Biol. Chem.* **1995**, *270* (23), 14031–14041.

[167] Fawaz, M. V.; Topper, M. E.; Firestine, S. M. The ATP-grasp enzymes. *Bioorg. Chem.* **2011**, *39* (5-6), 185–191.

[168] Benítez, D.; Medeiros, A.; Fiestas, L.; Panozzo-Zenere, E. A.; Maiwald, F.; Prousis, K. C.; Roussaki, M.; Calogeropoulou, T.; Detsi, A.; Jäger, T.; Šarlauskas, J.; Peterlin Mašič, L.; Kunick, C.; Labadie, G. R.; Flohé, L.; Comini, M. A. Identification of novel chemical scaffolds inhibiting trypanothione synthetase from pathogenic Trypanosomatids. *PLoS Neglected Trop. Dis.* **2016**, *10* (4), e0004617.

[169] Kunick, C.; Lauenroth, K.; Wieking, K.; Xie, X.; Schultz, C.; Gussio, R.; Zaharevitz, D. W.; Leost, M.; Meijer, L.; Weber, A.; Jørgensen, F. S.; Lemcke, T. Evaluation and comparison of 3D-QSAR CoMSIA models for CDK1, CDK5, and GSK-3 inhibition by paullones. *J. Med. Chem.* **2004**, *47* (1), 22–36.

[170] Maiwald, F.; Benítez, D.; Charquero, D.; Dar, M. A.; Erdmann, H.; Preu, L.; Koch, O.; Hölscher, C.; Loaëc, N.; Meijer, L.; Comini, M. A.; Kunick, C. 9- and 11-Substituted 4-azapaullones are potent and selective inhibitors of African trypanosoma. *Eur. J. Med. Chem.* **2014**, *83*, 274–283.

[171] Zhang, Z.; Li, Y.; Lin, B.; Schroeder, M.; Huang, B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **2011**, *27* (15), 2083–2088.

[172] Tan, K. P.; Varadarajan, R.; Madhusudhan, M. S. DEPTH: A web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res.* **2011**, *39* (Web Server issue), W242-8.

[173] Williams, M. A.; Goodfellow, J. M.; Thornton, J. M. Buried waters and internal cavities in monomeric proteins. *Protein Sci.* **1994**, *3* (8), 1224–1235.

[174] Dias, S. E. D.; Martins, A. M.; Nguyen, Q. T.; Gomes, A. J. P. GPU-based detection of protein cavities using Gaussian surfaces. *BMC Bioinf.* **2017**, *18* (1), 493.

[175] Voss, N. R.; Gerstein, M. 3V: Cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W555-62.

[176] Ho, B. K.; Gruswitz, F. HOLLOW: Generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct. Biol.* **2008**, *8*, 49.

[177] Kozlíková, B.; Sebestova, E.; Sustr, V.; Brezovsky, J.; Strnad, O.; Daniel, L.; Bednar, D.; Pavelka, A.; Manak, M.; Bezdeka, M.; Benes, P.; Kotry, M.; Gora, A.; Damborský, J.; Sochor, J. CAVER Analyst 1.0: Graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. *Bioinformatics* **2014**, *30* (18), 2684–2685.

[178] Hung, L. V.; Caprari, S.; Bizai, M.; Toti, D.; Polticelli, F. LIBRA: LIgand Binding site Recognition Application. *Bioinformatics* **2015**, *31* (24), 4020–4022.

[179] Brylinski, M.; Feinstein, W. P. eFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J. Comput.-Aided Mol. Des.* **2013**, *27* (6), 551–567.

[180] Pettit, F. K.; Bare, E.; Tsai, A.; Bowie, J. U. HotPatch: A statistical approach to finding biologically relevant features on protein surfaces. *J. Mol. Biol.* **2007**, *369* (3), 863–879.

[181] Kinjo, A. R.; Nakamura, H. Composite structural motifs of binding sites for delineating biological functions of proteins. *PLoS ONE* **2012**, *7* (2), e31437.

[182] Schneider, S.; Zacharias, M. Combining geometric pocket detection and desolvation properties to detect putative ligand binding sites on proteins. *J. Struct. Biol.* **2012**, *180* (3), 546–550.
[183] Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005**, *21* (9), 1908–1916.
[184] Buša, J.; Hayryan, S.; Hu, C.-K.; Skřivánek, J.; Wu, M.-C. CAVE: A package for detection and quantitative analysis of internal cavities in a system of overlapping balls: Application to proteins. *Comput. Phys. Commun.* **2010**, *181* (12), 2116–2125.
[185] Buša, J.; Hayryan, S.; Hu, C.-K.; Wu, M.-C. CAVE-CL: An OpenCL version of the package for detection and quantitative analysis of internal cavities in a system of overlapping balls: Application to proteins. *Comput. Phys. Commun.* **2015**, *190*, 224–227.
[186] Gao, M.; Skolnick, J. APoc: Large-scale identification of similar protein pockets. *Bioinformatics* **2013**, *29* (5), 597–604.
[187] Yuan, Y.; Pei, J.; Lai, L. Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.* **2013**, *19* (12), 2326–2333.
[188] Khangulov, V. *Protein Cavity Search*; John Hopkins University, Baltimore, USA, 2011.
[189] Coleman, R. G.; Sharp, K. A. Travel depth, a new shape descriptor for macromolecules: Application to ligand binding. *J. Mol. Biol.* **2006**, *362* (3), 441–458.
[190] Coleman, R. G.; Sharp, K. A. Protein pockets: Inventory, shape, and comparison. *J. Chem. Inf. Model.* **2010**, *50* (4), 589–603.
[191] Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.* **2010**, *50* (11), 2041–2052.
[192] Raunest, M.; Kandt, C. dxTuber: Detecting protein cavities, tunnels and clefts based on protein and solvent dynamics. *J. Mol. Graphics Modell.* **2011**, *29* (7), 895–905.
[193] Xu, D.; Zhang, Y. Generating triangulated macromolecular surfaces by Euclidean Distance Transform. *PLoS ONE* **2009**, *4* (12), e8140.
[194] Eyrisch, S.; Helms, V. Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.* **2007**, *50* (15), 3457–3464.
[195] Neudert, G.; Klebe, G. fconv: Format conversion, manipulation and feature computation of molecular data. *Bioinformatics* **2011**, *27* (7), 1021–1022.
[196] Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.
[197] Gaudreault, F.; Morency, L.-P.; Najmanovich, R. J. NRGsuite: a PyMOL plugin to perform docking simulations in real time using FlexAID. *Bioinformatics* **2015**, *31* (23), 3856–3858.
[198] Kawabata, T. Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* **2010**, *78* (5), 1195–1211.
[199] Oliveira, S. H. P.; Ferraz, F. A. N.; Honorato, R. V.; Xavier-Neto, J.; Sobreira, T. J. P.; de Oliveira, Paulo S L. KVFinder: Steered identification of protein cavities as a PyMOL plugin. *BMC Bioinf.* **2014**, *15*, 197.
[200] Till, M. S.; Ullmann, G. M. McVol - A program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. *J. Mol. Model.* **2010**, *16* (3), 419–429.

[201] Sehnal, D.; Svobodová Vařeková, R.; Berka, K.; Pravda, L.; Navrátilová, V.; Banáš, P.; Ionescu, C.-M.; Otyepka, M.; Koča, J. MOLE 2.0: Advanced approach for analysis of biomacromolecular channels. *J. Cheminf.* **2013**, *5* (1), 39.

[202] Zhu, H.; Pisabarro, M. T. MSPocket: An orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics* **2011**, *27* (3), 351–358.

[203] Decherchi, S.; Rocchia, W. A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. *PLoS ONE* **2013**, *8* (4), e59744.

[204] Kawabata, T.; Go, N. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins* **2007**, *68* (2), 516–529.

[205] Craig, I. R.; Pfleger, C.; Gohlke, H.; Essex, J. W.; Spiegel, K. Pocket-space maps to identify novel binding-site conformations in proteins. *J. Chem. Inf. Model.* **2011**, *51* (10), 2666–2679.

[206] Kalidas, Y.; Chandra, N. PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins. *J. Struct. Biol.* **2008**, *161* (1), 31–42.

[207] Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1*, 7.

[208] Czirják, G. PrinCCes: Continuity-based geometric decomposition and systematic visualization of the void repertoire of proteins. *J. Mol. Graphics Modell.* **2015**, *62*, 118–127.

[209] Laskowski, R. A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics* **1995**, *13* (5), 323-30, 307-8.

[210] Paramo, T.; East, A.; Garzón, D.; Ulmschneider, M. B.; Bond, P. J. Efficient characterization of protein cavities within molecular simulation trajectories: Trj_cavity. *J. Chem. Theory Comput.* **2014**, *10* (5), 2151–2164.

[211] Li, B.; Turuvekere, S.; Agrawal, M.; La, D.; Ramani, K.; Kihara, D. Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins* **2008**, *71* (2), 670–683.

[212] Mach, P.; Koehl, P. Geometric measures of large biomolecules: Surface, volume, and pockets. *J. Comput. Chem.* **2011**, *32* (14), 3023–3038.

[213] Rother, K.; Hildebrand, P. W.; Goede, A.; Gruening, B.; Preissner, R. Voronoia: Analyzing packing in protein structures. *Nucleic Acids Res.* **2009**, *37* (Database issue), 5.

[214] Ravindranath, P. A.; Sanner, M. F. AutoSite: An automated approach for pseudo-ligands prediction-from ligand-binding sites identification to predicting key ligand atoms. *Bioinformatics* **2016**, *32* (20), 3142–3149.

[215] Fukunishi, Y.; Nakamura, H. Prediction of ligand-binding sites of proteins by molecular docking calculation for a random ligand library. *Protein Sci.* **2011**, *20* (1), 95–106.

[216] Ghersi, D.; Sanchez, R. Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. *Proteins* **2009**, *74* (2), 417–424.

[217] Ghersi, D.; Sanchez, R. EasyMIFS and SiteHound: A toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **2009**, *25* (23), 3185–3186.

[218] Halgren, T. A. New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* **2007**, *69* (2), 146–148.

[219] Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49* (2), 377–389.

[220] Roy, A.; Zhang, Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* **2012**, *20* (6), 987–997.
[221] Lee, H. S.; Im, W. Ligand binding site detection by local structure alignment and its performance complementarity. *J. Chem. Inf. Model.* **2013**, *53* (9), 2462–2470.
[222] Xie, Z.-R.; Hwang, M.-J. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. *Bioinformatics* **2012**, *28* (12), 1579–1585.
[223] Krivák, R.; Hoksza, D. P2RANK: Knowledge-based ligand binding site prediction using aggregated local features. In *Algorithms for Computational Biology*; Dediu, A.-H., Hernández-Quiroz, F., Martín-Vide, C., Rosenblueth, D. A., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2015; pp 41–52.
[224] Tan, K. P.; Nguyen, T. B.; Patel, S.; Varadarajan, R.; Madhusudhan, M. S. Depth: A web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. *Nucleic Acids Res.* **2013**, *41* (Web Server issue), W314-21.
[225] Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J. Chem. Inf. Model.* **2014**, *54* (6), 1700–1716.
[226] Huang, W.; Wang, G.; Shen, Q.; Liu, X.; Lu, S.; Geng, L.; Huang, Z.; Zhang, J. ASBench: Benchmarking sets for allosteric discovery. *Bioinformatics* **2015**, *31* (15), 2598–2600.
[227] Schrödinger, L. L.C. *The PyMOL Molecular Graphics System, Version 1.8*.
[228] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.
[229] Kiefer, F.; Arnold, K.; Künzli, M.; Bordoli, L.; Schwede, T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* **2009**, *37* (Database issue), D387-92.
[230] Akbar, R.; Helms, V. ALLO: A tool to discriminate and prioritize allosteric pockets. *Chem. Biol. Drug Des.* **2018**, *91* (4), 845–853.
[231] Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33-8, 27-8.
[232] Masuya, M.; Doi, J. Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. *J. Mol. Graphics* **1995**, *13* (6), 331–336.
[233] Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32* (Web Server issue), W665-7.
[234] PDB2PQR. https://sourceforge.net/projects/pdb2pqr/ (accessed 01/27/2018).
[235] Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234* (3), 779–815.

[236] Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. *Amber 14*, 2014.

[237] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.

[238] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72* (2), 171–179.

[239] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

[240] Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. Overview of the CCP4 suite and current developments. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2011**, *67* (Pt 4), 235–242.

[241] *Schrödinger Suite 2017-2 Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2016; Impact, Schrödinger, LLC, New York, NY, 2016; Prime, Schrödinger, LLC*; Schrödinger, LLC: New York, NY, 2017.

[242] *SiteMap, Schrödinger, LLC, New York, NY*; Schrödinger, LLC: New York, NY, 2017.

[243] Lee, H. S.; Im, W. G-LoSA for prediction of protein-ligand binding sites and structures. *Methods Mol. Biol.* **2017**, *1611*, 97–108.

[244] Altschul, S. F. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–3402.

[245] Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinf.* **2008**, *9*, 40.

[246] Schneider, R.; Daruvar, A. de; Sander, C. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.* **1997**, *25* (1), 226–230.

[247] HSSP. ftp://ftp.cmbi.ru.nl/pub/molbio/data/hssp/ (accessed 01/27/2018).

[248] ConSeq. https://conseq.tau.ac.il/overviewnextversion.htm (accessed 01/27/2018).

[249] Capra, J. A.; Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **2007**, *23* (15), 1875–1882.

[250] ConCavity (conservation). http://compbio.cs.princeton.edu/conservation/ (accessed 01/27/2018).

[251] Yu, J.; Zhou, Y.; Tanaka, I.; Yao, M. Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics* **2010**, *26* (1), 46–52.

[252] Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405* (2), 442–451.

[253] López, G.; Ezkurdia, I.; Tress, M. L. Assessment of ligand binding residue predictions in CASP8. *Proteins* **2009**, *77 Suppl 9*, 138–146.

[254] Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.* **2012**, *52* (2), 360–372.

[255] Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26* (2), 283–291.

[256] Laskowski, R. A.; Rullmannn, J. A.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **1996**, *8* (4), 477–486.

[257] *Molecular Operating Environment (MOE)*; Chemical Computing Group ULC: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.

[258] Wang, G.; Dunbrack, R. L., JR. PISCES: A protein sequence culling server. *Bioinformatics* **2003**, *19* (12), 1589–1591.

[259] Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26* (19), 2460–2461.

[260] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

[261] Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer, 2007.

[262] Gfeller, D.; Michielin, O.; Zoete, V. SwissSidechain: A molecular and structural database of non-natural sidechains. *Nucleic Acids Res.* **2013**, *41* (Database issue), 32.

[263] Dunbrack, R. L. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **2002**, *12* (4), 431–440.

[264] *ROCS*; OpenEye Scientific Software: Santa Fe, NM.

[265] Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Favia, A. D.; Thornton, J. M. On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins* **2010**, *78* (5), 1120–1136.

[266] Hoffmann, B.; Zaslavskiy, M.; Vert, J.-P.; Stoven, V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: Application to ligand prediction. *BMC Bioinf.* **2010**, *11*, 99.

[267] Barelier, S.; Sterling, T.; O'Meara, M. J.; Shoichet, B. K. The recognition of identical ligands by unrelated proteins. *ACS Chem. Biol.* **2015**, *10* (12), 2772–2784.

[268] Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding protein-ligand interaction patterns in fingerprints and graphs. *J. Chem. Inf. Model.* **2013**, *53* (3), 623–637.

[269] Chartier, M.; Najmanovich, R. J. Detection of binding site molecular interaction field similarities. *J. Chem. Inf. Model.* **2015**, *55* (8), 1600–1615.

[270] Krotzky, T.; Grunwald, C.; Egerland, U.; Klebe, G. Large-scale mining for similar protein binding pockets: With RAPMAD retrieval on the fly becomes real. *J. Chem. Inf. Model.* **2015**, *55* (1), 165–179.

[271] Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. SiteEngines: Recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), 41.

[272] Batista, J.; Hawkins, P. C. D.; Tolbert, R.; Geballe, M. T. SiteHopper - a unique tool for binding site comparison. *J. Cheminf.* **2014**, *6* (Suppl 1), P57.

[273] Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15* (6), 359.

[274] DeLong, E. R.; DeLong, D. M.; Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44* (3), 837–845.

[275] Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.* **2011**, *12*, 77.

[276] R Core Team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.

[277] Koch, O. Secbase: A novel tool to correlate secondary structure elements with ligand binding. Dissertation, Philipps-Universität Marburg, Marburg, 2008.

[278] Koch, O.; Klebe, G. Turns revisited: A uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. *Proteins* **2009**, *74* (2), 353–367.

[279] Ultsch, Alfred, Mörchen, Fabian. *ESOM-Maps: Tools for clustering, visualization, and classification with Emergent SOM* Technical Report No. 46; University of Marburg, Germany, 2005.

[280] McDonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238* (5), 777–793.

[281] Zacharias, J. Assignment of local protein structure with different strategies. Dissertation, Freie Universität Berlin, Berlin, 2014.

[282] Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.

[283] Boobbyer, D. N.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.* **1989**, *32* (5), 1083–1094.

[284] Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849–857.

[285] Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94* (26), 8897–8909.

[286] Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6* (6), 1333–1337.

[287] Fabiola, F.; Bertram, R.; Korostelev, A.; Chapman, M. S. An improved hydrogen bond potential: Impact on medium resolution protein structures. *Protein Sci.* **2002**, *11* (6), 1415–1423.

[288] Brünger, A. T. *X-PLOR, Version 3.1: A system for X-ray crystallography and NMR*; Yale University Press: New Haven, 1992.

[289] Ultsch, A. Maps for visualization of high-dimensional data spaces. *WSOM: Kyushu, Japan* **2003**, 225–230.

[290] Batschelet, E. *Circular statistics in biology*; Academic Press: San Diego, London, Boston, 1981.

[291] Chou, P. Y.; Fasman, G. D. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry* **1974**, *13* (2), 211–222.

[292] Wilmot, C. M.; Thornton, J. M. Analysis and prediction of the different types of β-turn in proteins. *J. Mol. Biol.* **1988**, *203* (1), 221–232.

[293] Carugo, O.; Argos, P. Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. *Proteins* **1998**, *31* (2), 201–213.

[294] Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9* (9), 3878–3888.

[295] Le Grand, S.; Götz, A. W.; Walker, R. C. SPFP: Speed without compromise - A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.* **2013**, *184* (2), 374–380.

[296] Meagher, K. L.; Redman, L. T.; Carlson, H. A. Development of polyphosphate parameters for use with the AMBER force field. *J. Comput. Chem.* **2003**, *24* (9), 1016–1025.

[297] Walker, R. C.; Crowley, M. F.; Case, D. A. The implementation of a fast and accurate QM/MM potential method in Amber. *J. Comput. Chem.* **2008**, *29* (7), 1019–1031.

[298] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. L. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713.

[299] Joung, I. S.; Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041.

[300] Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.

[301] Shapiro, J.; Brutlag, D. FoldMiner and LOCK2: Protein structure comparison and motif discovery on the web. *Nucleic Acids Res.* **2004**, *32* (Web Server issue), W536-41.

[302] Kooistra, A. J.; Kanev, G. K.; van Linden, O. P. J.; Leurs, R.; Esch, I. J. P. de; Graaf, C. de. KLIFS: A structural kinase-ligand interaction database. *Nucleic Acids Res.* **2016**, *44* (D1), D365-71.

[303] Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* **2014**, *42* (Database issue), D1083-90.

[304] Morgan, H. L. The generation of a unique machine description for chemical structures - A technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113.
[305] Landrum, G. *RDKit*, 2016.
[306] Casari, G.; Sander, C.; Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Mol. Biol.* **1995**, *2* (2), 171–178.
[307] Brown, D.; Superti-Furga, G. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today* **2003**, *8* (23), 1067–1077.
[308] Ringe, D. What makes a binding site a binding site? *Curr. Opin. Struct. Biol.* **1995**, *5* (6), 825–829.
[309] Levitt, D. G.; Banaszak, L. J. POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graphics* **1992**, *10* (4), 229–234.
[310] Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* **1996**, *22* (4), 469–483.
[311] Buša, J.; Hayryan, S.; Hu, C.-K.; Skřivánek, J.; Wu, M.-C. Enveloping triangulation method for detecting internal cavities in proteins and algorithm for computing their surface areas and volumes. *J. Comput. Chem.* **2009**, *30* (3), 346–357.
[312] Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38* (3), 305–320.
[313] Stahl, M.; Taroni, C.; Schneider, G. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng., Des. Sel.* **2000**, *13* (2), 83–88.
[314] Edelsbrunner, H. The union of balls and its dual shape. *Discrete Comput. Geom.* [Online] **1995**, *13* (3), 415–440. https://doi.org/10.1007/BF02574053.
[315] Kim, D.-S.; Seo, J.; Kim, D.; Ryu, J.; Cho, C.-H. Three-dimensional β shapes. *Comput. Aided Design* **2006**, *38* (11), 1179–1191.
[316] Kim, D.-S.; Cho, Y.; Sugihara, K.; Ryu, J.; Kim, D. Three-dimensional β-shapes and β-complexes via quasi-triangulation. *Comput. Aided Design* **2010**, *42* (10), 911–929.
[317] Lee, B.; Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **1971**, *55* (3), 379–400.
[318] Edelsbrunner, H.; Mücke, E. P. Three-dimensional α shapes. *ACM Trans. Graph.* **1994**, *13* (1), 43–72.
[319] An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics* **2005**, *4* (6), 752–761.
[320] Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288.
[321] Ewing, T. J.A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15* (5), 411–428.
[322] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46* (1-3), 3–26.

[323] Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14* (3), 251–264.
[324] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19* (14), 1639–1662.
[325] Fukunishi, Y.; Mikami, Y.; Nakamura, H. Similarities among receptor pockets and among compounds: Analysis and application to *in silico* ligand screening. *J. Mol. Graphics Modell.* **2005**, *24* (1), 34–45.
[326] Fukunishi, Y.; Sugihara, Y.; Mikami, Y.; Sakai, K.; Kusudo, H.; Nakamura, H. Advanced *in silico* drug screening to achieve high hit ratio. *Synthesiology* **2009**, *2* (1), 64–72.
[327] Xie, Z.-R.; Hwang, M.-J. An interaction-motif-based scoring function for protein-ligand docking. *BMC Bioinf.* **2010**, *11*, 298.
[328] Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; Fabritiis, G. de. DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017**, *33* (19), 3036–3042.
[329] Krivák, R.; Hoksza, D. P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminf.* **2018**, *10* (1), 39.
[330] Hubbard, S. J.; Thornton, J. M. *NACCESS*; Department of Biochemistry and Molecular Biology, University College London, 1993.
[331] Shrake, A.; Rupley, J. A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **1973**, *79* (2), 351–371.
[332] Brylinski, M.; Skolnick, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (1), 129–134.
[333] Yang, J.; Roy, A.; Zhang, Y. BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **2013**, *41* (Database issue), D1096-103.
[334] Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292* (2), 195–202.
[335] Benchmark Datasets used in metaPocket 2.0. http://projects.biotec.tu-dresden.de/metapocket/benchmark.php (accessed December 15, 2018).
[336] Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7* (9), 1884–1897.
[337] Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267* (3), 727–748.
[338] Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47* (12), 2977–2980.
[339] Puvanendrampillai, D.; Mitchell, J. B. O. Protein Ligand Database (PLD): Additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics* **2003**, *19* (14), 1856–1857.

[340] Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25* (1), 71–75.

[341] Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50* (4), 726–741.

[342] Dessailly, B. H.; Lensink, M. F.; Orengo, C. A.; Wodak, S. J. LigASite - a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.* **2008**, *36* (Database issue), D667-73.

[343] Laskowski, R. A.; Jabłońska, J.; Pravda, L.; Vařeková, R. S.; Thornton, J. M. PDBsum: Structural summaries of PDB entries. *Protein Sci.* **2018**, *27* (1), 129–134.

[344] Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56* (2), 235–249.

[345] Krasowski, A.; Muthas, D.; Sarkar, A.; Schmitt, S.; Brenk, R. DrugPred: A structure-based approach to predict protein druggability developed using an extensive nonredundant data set. *J. Chem. Inf. Model.* **2011**, *51* (11), 2829–2842.

[346] Roche, O.; Kiyama, R.; Brooks, C. L. Ligand-protein database: Linking protein-ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44* (22), 3592–3598.

[347] Spronk, C. A. E. M.; Nabuurs, S. B.; Krieger, E.; Vriend, G.; Vuister, G. W. Validation of protein structures derived by NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **2004**, *45* (3-4), 315–337.

[348] Thangudu, R. R.; Tyagi, M.; Shoemaker, B. A.; Bryant, S. H.; Panchenko, A. R.; Madej, T. Knowledge-based annotation of small molecule binding sites in proteins. *BMC Bioinf.* **2010**, *11*, 365.

[349] Joosten, R. P.; te Beek, T. A. H.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W. W.; Schneider, R.; Sander, C.; Vriend, G. A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **2011**, *39* (Database issue), D411-9.

[350] Wagner, J. R.; Lee, C. T.; Durrant, J. D.; Malmstrom, R. D.; Feher, V. A.; Amaro, R. E. Emerging computational methods for the rational discovery of allosteric drugs. *Chem. Rev.* **2016**, *116* (11), 6370–6390.

[351] Jalencas, X.; Mestres, J. Identification of similar binding sites to detect distant polypharmacology. *Mol. Inf.* **2013**, *32* (11-12), 976–990.

[352] Štular, T.; Lešnik, S.; Rožman, K.; Schink, J.; Zdouc, M.; Ghysels, A.; Liu, F.; Aldrich, C. C.; Haupt, V. J.; Salentin, S.; Daminelli, S.; Schroeder, M.; Langer, T.; Gobec, S.; Janežič, D.; Konc, J. Discovery of *Mycobacterium tuberculosis* InhA inhibitors by binding sites comparison and ligands prediction. *J. Med. Chem.* **2016**, *59* (24), 11069–11078.

[353] Yang, L.; Chen, J.; Shi, L.; Hudock, M. P.; Wang, K.; He, L. Identifying unexpected therapeutic targets via chemical-protein interactome. *PLoS ONE* **2010**, *5* (3), e9568.

[354] Al-Gharabli, S. I.; Shah, S. T. A.; Weik, S.; Schmidt, M. F.; Mesters, J. R.; Kuhn, D.; Klebe, G.; Hilgenfeld, R.; Rademann, J. An efficient method for the synthesis of peptide aldehyde libraries employed in the discovery of reversible SARS coronavirus main protease (SARS-CoV Mpro) inhibitors. *ChemBioChem* **2006**, *7* (7), 1048–1055.

[355] Kuhn, D.; Weskamp, N.; Schmitt, S.; Hüllermeier, E.; Klebe, G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **2006**, *359* (4), 1023–1044.

[356] Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: New pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, *47* (3), 550–557.

[357] Sturm, N.; Desaphy, J.; Quinn, R. J.; Rognan, D.; Kellenberger, E. Structural insights into the molecular basis of the ligand promiscuity. *J. Chem. Inf. Model.* **2012**, *52* (9), 2410–2421.

[358] Kurbatova, N.; Chartier, M.; Zylber, M. I.; Najmanovich, R. J. IsoCleft Finder - A web-based tool for the detection and analysis of protein binding-site geometric and chemical similarities. *F1000Res.* **2013**, *2*, 117.

[359] Brylinski, M. eMatchSite: Sequence order-independent structure alignments of ligand binding pockets in protein models. *PLoS Comput. Biol.* **2014**, *10* (9), e1003829.

[360] Larocque, M.; Chenard, T.; Najmanovich, R. J. A curated *C. difficile* strain 630 metabolic network: prediction of essential targets and inhibitors. *BMC Syst. Biol.* **2014**, *8*, 117.

[361] Schirris, T. J. J.; Ritschel, T.; Herma Renkema, G.; Willems, P. H. G. M.; Smeitink, J. A. M.; Russel, F. G. M. Mitochondrial ADP/ATP exchange inhibition: A novel off-target mechanism underlying ibipinabant-induced myotoxicity. *Sci. Rep.* **2015**, *5*, 14533.

[362] McGuire, R.; Verhoeven, S.; Vass, M.; Vriend, G.; de Esch, Iwan J P; Lusher, S. J.; Leurs, R.; Ridder, L.; Kooistra, A. J.; Ritschel, T.; Graaf, C. de. 3D-e-Chem-VM: Structural cheminformatics research infrastructure in a freely available virtual machine. *J. Chem. Inf. Model.* **2017**, *57* (2), 115–121.

[363] Schirris, T. J. J.; Renkema, G. H.; Ritschel, T.; Voermans, N. C.; Bilos, A.; van Engelen, B. G. M.; Brandt, U.; Koopman, W. J. H.; Beyrath, J. D.; Rodenburg, R. J.; Willems, P. H. G. M.; Smeitink, J. A. M.; Russel, F. G. M. Statin-induced myopathy is associated with mitochondrial complex III inhibition. *Cell Metab.* **2015**, *22* (3), 399–407.

[364] Anand, P.; Sankaran, S.; Mukherjee, S.; Yeturu, K.; Laskowski, R. A.; Bhardwaj, A.; Bhagavat, R.; Brahmachari, S. K.; Chandra, N. Structural annotation of Mycobacterium tuberculosis proteome. *PLoS ONE* **2011**, *6* (10), e27044.

[365] Moller-Acuna, P.; Contreras-Riquelme, J. S.; Rojas-Fuentes, C.; Nunez-Vivanco, G.; Alzate-Morales, J.; Iturriaga-Vasquez, P.; Arias, H. R.; Reyes-Parada, M. Similarities between the binding sites of SB-206553 at serotonin type 2 and α7 acetylcholine nicotinic receptors: Rationale for its polypharmacological profile. *PLoS ONE* **2015**, *10* (8), e0134444.

[366] Fierro, A.; Montecinos, A.; Gmez-Molina, C.; Nez, G.; Aldeco, M.; Edmondson, D. E.; Vilches Herrera, M.; Luhr, S.; Iturriaga-Vasquez, P.; Reyes-Parada, M. Similarities between the binding sites of monoamine oxidase (MAO) from different species - Is zebrafish a useful model for the discovery of novel MAO inhibitors? In *An Integrated View of the Molecular Recognition and Toxinology - From Analytical Procedures to Biomedical Applications*; Radis-Baptista, G., Ed.; InTech, 2013.

[367] Konc, J.; Hodoscek, M.; Ogrizek, M.; Trykowska Konc, J.; Janežič, D. Structure-based function prediction of uncharacterized protein using binding sites comparison. *PLoS Comput. Biol.* **2013**, *9* (11), e1003341.

[368] Kakisaka, M.; Sasaki, Y.; Yamada, K.; Kondoh, Y.; Hikono, H.; Osada, H.; Tomii, K.; Saito, T.; Aida, Y. A novel antiviral target structure involved in the RNA binding, dimerization, and nuclear export functions of the influenza A virus nucleoprotein. *PLoS Pathog.* **2015**, *11* (7), e1005062.

[369] Krotzky, T.; Fober, T.; Hüllermeier, E.; Klebe, G. Extended graph-based models for enhanced similarity search in Cavbase. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2014**, *11* (5), 878–890.

[370] Defranchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS ONE* **2010**, *5* (8), e12214.

[371] Keren-Kaplan, T.; Attali, I.; Estrin, M.; Kuo, L. S.; Farkash, E.; Jerabek-Willemsen, M.; Blutraich, N.; Artzi, S.; Peri, A.; Freed, E. O.; Wolfson, H. J.; Prag, G. Structure-based *in silico* identification of ubiquitin-binding domains provides insights into the ALIX-V:ubiquitin complex and retrovirus budding. *EMBO J.* **2013**, *32* (4), 538–551.

[372] Meyers, J.; Brown, N.; Blagg, J. Mapping the 3D structures of small molecule binding sites. *J. Cheminf.* **2016**, *8* (1), 235.

[373] Xie, L.; Evangelidis, T.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: Weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS Comput. Biol.* **2011**, *7* (4), e1002037.

[374] Niu, M.; Hu, J.; Wu, S.; Zhang, X.; Xu, H.; Zhang, Y.; Zhang, J.; Yang, Y. Structural bioinformatics-based identification of EGFR inhibitor gefitinib as a putative lead compound for BACE. *Chem. Biol. Drug Des.* **2014**, *83* (1), 81–88.

[375] Kooistra, A. J.; Leurs, R.; de Esch, Iwan J P; Graaf, C. de. Structure-based prediction of G-protein-coupled receptor ligand function: A β-adrenoceptor case study. *J. Chem. Inf. Model.* **2015**, *55* (5), 1045–1061.

[376] Yang, Y.; Li, G.; Zhao, D.; Yu, H.; Zheng, X.; Peng, X.; Zhang, X.; Fu, T.; Hu, X.; Niu, M.; Ji, X.; Zou, L.; Wang, J. Computational discovery and experimental verification of tyrosine kinase inhibitor pazopanib for the reversal of memory and cognitive deficits in rat model neurodegeneration. *Chem. Sci.* **2015**, *6* (5), 2812–2821.

[377] McLachlan, A. D. Repeating sequences and gene duplication in proteins. *J. Mol. Biol.* **1972**, *64* (2), 417–437.

[378] Karlin, S.; Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U. S. A.* **1990**, *87* (6), 2264–2268.

[379] *Shape Toolkit*; OpenEye Scientific Software: Santa Fe, NM.

[380] *Spicoli Toolkit*; OpenEye Scientific Software: Santa Fe, NM.

[381] Clark, M.; Cramer, R. D.; van Opdenbosch, N. Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10* (8), 982–1012.

[382] Fligner, M. A.; Verducci, J. S.; Blower, P. E. A modification of the Jaccard–Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics* **2002**, *44* (2), 110–119.
[383] Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45* (1), 160–169.
[384] Nakamura, T.; Tomii, K. Effects of the difference in similarity measures on the comparison of ligand-binding pockets using a reduced vector representation of pockets. *Biophys. Physicobiol.* **2016**, *13*, 139–147.
[385] Stierand, K.; Rarey, M. Drawing the PDB: Protein-ligand complexes in two dimensions. *ACS Med. Chem. Lett.* **2010**, *1* (9), 540–545.
[386] Kinnings, S. L.; Jackson, R. M. Binding site similarity analysis for the functional classification of the protein kinase family. *J. Chem. Inf. Model.* **2009**, *49* (2), 318–329.
[387] Zhao, Z.; Xie, L.; Xie, L.; Bourne, P. E. Delineation of polypharmacology across the human structural kinome using a functional site interaction fingerprint approach. *J. Med. Chem.* **2016**, *59* (9), 4326–4341.
[388] Volkamer, A.; Eid, S.; Turk, S.; Rippmann, F.; Fulle, S. Identification and visualization of kinase-specific subpockets. *J. Chem. Inf. Model.* **2016**, *56* (2), 335–346.
[389] Chung, C.-W.; Dean, A. W.; Woolven, J. M.; Bamborough, P. Fragment-based discovery of bromodomain inhibitors part 1: Inhibitor binding modes and implications for lead discovery. *J. Med. Chem.* **2012**, *55* (2), 576–586.
[390] Yang, Y.; Chen, H.; Nilsson, I.; Muresan, S.; Engkvist, O. Investigation of the relationship between topology and selectivity for druglike molecules. *J. Med. Chem.* **2010**, *53* (21), 7709–7714.
[391] Peters, J.-U.; Hert, J.; Bissantz, C.; Hillebrecht, A.; Gerebtzoff, G.; Bendels, S.; Tillier, F.; Migeon, J.; Fischer, H.; Guba, W.; Kansy, M. Can we discover pharmacological promiscuity early in the drug discovery process? *Drug Discovery Today* **2012**, *17* (7-8), 325–335.
[392] Young, R. J.; Green, D. V. S.; Luscombe, C. N.; Hill, A. P. Getting physical in drug discovery II: the impact of chromatographic hydrophobicity measurements and aromaticity. *Drug Discovery Today* **2011**, *16* (17-18), 822–830.
[393] Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486* (7403), 361–367.
[394] Ghuman, J.; Zunszain, P. A.; Petitpas, I.; Bhattacharya, A. A.; Otagiri, M.; Curry, S. Structural basis of the drug-binding specificity of human serum albumin. *J. Mol. Biol.* **2005**, *353* (1), 38–52.
[395] Khan, H.; Harris, R. J.; Barna, T.; Craig, D. H.; Bruce, N. C.; Munro, A. W.; Moody, P. C. E.; Scrutton, N. S. Kinetic and structural basis of reactivity of pentaerythritol tetranitrate reductase with NADPH, 2-cyclohexenone, nitroesters, and nitroaromatic explosives. *J. Biol. Chem.* **2002**, *277* (24), 21906–21912.
[396] Funder, J. W. The promiscuous receptor: A case for the guardian enzyme. *Cardiovasc. Res.* **1995**, *30* (2), 177–180.

[397] Carbonell, P.; Faulon, J.-L. Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics* **2010**, *26* (16), 2012–2019.
[398] Schreiber, G.; Keating, A. E. Protein binding specificity versus promiscuity. *Curr. Opin. Struct. Biol.* **2011**, *21* (1), 50–61.
[399] Baldi, P.; Nasr, R. When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *J. Chem. Inf. Model.* **2010**, *50* (7), 1205–1222.
[400] Jasial, S.; Hu, Y.; Vogt, M.; Bajorath, J. Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Res.* **2016**, *5*.
[401] Vogt, M.; Bajorath, J. Modeling Tanimoto similarity value distributions and predicting search results. *Mol. Inf.* **2017**, *36* (7).
[402] Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold hopping using two-dimensional fingerprints: True potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. *J. Med. Chem.* **2010**, *53* (15), 5707–5715.
[403] Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking molecular similarity: Comparing compounds on the basis of biological activity. *ACS Chem. Biol.* **2012**, *7* (8), 1399–1409.
[404] Youden, W. J. Index for rating diagnostic tests. *Cancer* **1950**, *3* (1), 32–35.
[405] van Rijsbergen, C. J. *Information retrieval*; Butterworth: London, 1979.
[406] Biggerstaff, B. J. Comparing diagnostic tests: A simple graphic using likelihood ratios. *Stat. Med.* **2000**, *19* (5), 649–663.
[407] Blakeley, D. D. Noninvasive carotid artery testing: A meta-analytic review. *Ann. Intern. Med.* **1995**, *122* (5), 360.
[408] Stark, A.; Sunyaev, S.; Russell, R. B. A model for statistical significance of local similarities in structure. *J. Mol. Biol.* **2003**, *326* (5), 1307–1316.
[409] Nabuurs, S. B.; Spronk, C. A. E. M.; Vuister, G. W.; Vriend, G. Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. *PLoS Comput. Biol.* **2006**, *2* (2), e9.
[410] Chen, Y. W. Solution solution: Using NMR models for molecular replacement. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2001**, *57* (10), 1457–1461.
[411] Chaudhury, S.; Gray, J. J. Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. *J. Mol. Biol.* **2008**, *381* (4), 1068–1087.
[412] Slynko, I.; Da Silva, F.; Bret, G.; Rognan, D. Docking pose selection by interaction pattern graph similarity: application to the D3R grand challenge 2015. *J. Comput.-Aided Mol. Des.* **2016**, *30* (9), 669–683.
[413] Krotzky, T. Methods for the efficient comparison of protein binding sites and for the assessment of protein-ligand complexes. Dissertation, Philipps-Universität Marburg, Marburg, 2015.
[414] Chen, Y.-C.; Tolbert, R.; Aronov, A. M.; McGaughey, G.; Walters, W. P.; Meireles, L. Prediction of protein pairs sharing common active ligands using protein sequence, structure, and ligand similarity. *J. Chem. Inf. Model.* **2016**, *56* (9), 1734–1745.

[415] VIB / UGent, Bioinformatics & Evolutionary Genomics, Belgium. DrawVenn: Calculate and draw custom Venn diagrams. http://bioinformatics.psb.ugent.be/webtools/Venn/.
[416] Koch, O.; Cole, J. C.; Block, P.; Klebe, G. Secbase: Database module to retrieve secondary structure elements with ligand binding motifs. *J. Chem. Inf. Model.* **2009**, *49* (10), 2388–2402.
[417] Bruno, I. J.; Cole, J. C.; Lommerse, J. P.M.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. IsoStar: A library of information about nonbonded interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11* (6), 525–537.
[418] Emberly, E. G.; Wingreen, N. S.; Tang, C. Designability of α-helical proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (17), 11163–11168.
[419] Liang, S.; Grishin, N. V. Side-chain modeling with an optimized scoring function. *Protein Sci.* **2002**, *11* (2), 322–331.
[420] Rader, A. J.; Hespenheide, B. M.; Kuhn, L. A.; Thorpe, M. F. Protein unfolding: Rigidity lost. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (6), 3540–3545.
[421] Livesay, D. R.; Huynh, D. H.; Dallakyan, S.; Jacobs, D. J. Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family. *Chem. Cent. J.* **2008**, *2*, 17.
[422] Kyngas, J.; Valjakka, J. Unreliability of the Chou-Fasman parameters in predicting protein secondary structure. *Protein Eng., Des. Sel.* **1998**, *11* (5), 345–348.
[423] Ramachandran, G. N.; Chandrasekaran, R. Conformation of peptide chains containing both L- & D-residues. I. Helical structures with alternating L- & D-residues with special reference to the LD-ribbon & the LD-helices. *Indian J. Biochem. Biophys.* **1972**, *9* (1), 1–11.
[424] Smith, L. J.; Bolin, K. A.; Schwalbe, H.; MacArthur, M. W.; Thornton, J. M.; Dobson, C. M. Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. *J. Mol. Biol.* **1996**, *255* (3), 494–506.
[425] Enkhbayar, P.; Boldgiv, B.; Matsushima, N. ω-Helices in proteins. *Protein J.* **2010**, *29* (4), 242–249.
[426] van der Kant, R.; Vriend, G. α-bulges in G protein-coupled receptors. *Int. J. Mol. Sci.* **2014**, *15* (5), 7841–7864.
[427] Adzhubei, A. A.; Sternberg, M. J. E. Left-handed polyproline II helices commonly occur in globular proteins. *J. Mol. Biol.* **1993**, *229* (2), 472–493.
[428] Bochicchio, B.; Tamburro, A. M. Polyproline II structure in proteins: Identification by chiroptical spectroscopies, stability, and functions. *Chirality* **2002**, *14* (10), 782–792.
[429] Marcelino, A. M. C.; Gierasch, L. M. Roles of β-turns in protein folding: From peptide models to protein engineering. *Biopolymers* **2008**, *89* (5), 380–391.
[430] Rose, G. D.; Gierasch, L. M.; Smith, J. A. Turns in peptides and proteins. *Adv. Protein Chem.* **1985**, *37*, 1–109.
[431] Efimov, A. V. Structure of coiled β-β-hairpins and β-β-corners. *FEBS Lett.* **1991**, *284* (2), 288–292.
[432] Dasgupta, B.; Chakrabarti, P. π-Turns: Types, systematics and the context of their occurrence in protein structures. *BMC Struct. Biol.* **2008**, *8*, 39.
[433] Aurora, R.; Rose, G. D. Helix capping. *Protein Sci.* **1998**, *7* (1), 21–38.

[434] Segura, J.; Oliva, B.; Fernandez-Fuentes, N. CAPS-DB: A structural classification of helix-capping motifs. *Nucleic Acids Res.* **2012**, *40* (Database issue), D479-85.

[435] Newell, N. E. Mapping side chain interactions at protein helix termini. *BMC Bioinf.* **2015**, *16*, 231.

[436] Schellman, C. The αL conformation at the ends of helices. *Protein Folding: Conference Proceedings* **1980**, 53–61.

[437] Milner-White, E. J. Recurring loop motif in proteins that occurs in right-handed and left-handed forms. *J. Mol. Biol.* **1988**, *199* (3), 503–511.

[438] Jiménez, A. I.; Ballano, G.; Cativiela, C. First observation of two consecutive γ turns in a crystalline linear dipeptide. *Angew. Chem. Int. Ed.* **2005**, *44* (3), 396–399.

[439] Tsunemi, M.; Matsuura, Y.; Sakakibara, S.; Katsube, Y. Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 Å resolution. *Biochemistry* **1996**, *35* (36), 11570–11576.

[440] Yang, M.; Culhane, J. C.; Szewczuk, L. M.; Gocke, C. B.; Brautigam, C. A.; Tomchick, D. R.; Machius, M.; Cole, P. A.; Yu, H. Structural basis of histone demethylation by LSD1 revealed by suicide inactivation. *Nat. Struct. Mol. Biol.* **2007**, *14* (6), 535–539.

[441] Smith, K.; Nadeau, K.; Bradley, M.; Walsh, C. T.; Fairlamb, A. H. Purification of glutathionylspermidine and trypanothione synthetases from *Crithidia fasciculata*. *Protein Sci.* **1992**, *1* (7), 874–883.

[442] Oza, S. L.; Ariyanayagam, M. R.; Fairlamb, A. H. Characterization of recombinant glutathionylspermidine synthetase/amidase from *Crithidia fasciculata*. *Biochem. J.* **2002**, *364* (Pt 3), 679–686.

[443] Oza, S. L.; Wyllie, S.; Fairlamb, A. H. Mapping the functional synthetase domain of trypanothione synthetase from *Leishmania major*. *Mol. Biochem. Parasitol.* **2006**, *149* (1), 117–120.

[444] Orban, O. C. F.; Korn, R. S.; Benítez, D.; Medeiros, A.; Preu, L.; Loaëc, N.; Meijer, L.; Koch, O.; Comini, M. A.; Kunick, C. 5-Substituted 3-chlorokenpaullone derivatives are potent inhibitors of *Trypanosoma brucei* bloodstream forms. *Bioorg. Med. Chem.* **2016**, *24* (16), 3790–3800.

[445] Koch, O.; Jäger, T.; Flohé, L.; Selzer, P. M. Inhibition of trypanothione synthetase as a therapeutic concept. In *Trypanosomatid Diseases*; Jäger, T., Koch, O., Flohé, L., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2013; pp 429–443.

[446] Bain, J.; Plater, L.; Elliott, M.; Shpiro, N.; Hastie, C. J.; McLauchlan, H.; Klevernic, I.; Arthur, J. S. C.; Alessi, D. R.; Cohen, P. The selectivity of protein kinase inhibitors: A further update. *Biochem. J.* **2007**, *408* (3), 297–315.

[447] EMBL-EBI. Kinase SARfari. https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari (accessed December 17, 2018).

[448] Boateng, N. Computerbasierte Suche nach neuen Trypanothionsynthetase-Inhibitoren. Bachelor thesis, TU Dortmund University, Dortmund, 2014.

[449] Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298* (5600), 1912–1934.

[450] Eid, S.; Turk, S.; Volkamer, A.; Rippmann, F.; Fulle, S. KinMap: A web-based tool for interactive navigation through human kinome data. *BMC Bioinf.* **2017**, *18* (1), 16.

[451] Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: Repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, *5* (7), e1000423.

[452] Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011**, *29* (11), 1046–1051.

[453] Byrne, D. P.; Vonderach, M.; Ferries, S.; Brownridge, P. J.; Eyers, C. E.; Eyers, P. A. cAMP-dependent protein kinase (PKA) complexes probed by complementary differential scanning fluorimetry and ion mobility-mass spectrometry. *Biochem. J.* **2016**, *473* (19), 3159–3175.

[454] Esser, L.; Wang, C. R.; Hosaka, M.; Smagula, C. S.; Südhof, T. C.; Deisenhofer, J. Synapsin I is structurally similar to ATP-utilizing enzymes. *EMBO J.* **1998**, *17* (4), 977–984.

[455] Zhu, T.; Cao, S.; Su, P.-C.; Patel, R.; Shah, D.; Chokshi, H. B.; Szukala, R.; Johnson, M. E.; Hevener, K. E. Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis. *J. Med. Chem.* **2013**, *56* (17), 6560–6572.

[456] Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P. J.; Whitmore, A. V.; Zimmer, S.; Young, M. P.; Jenkins, J. L.; Glick, M.; Glen, R. C.; Bender, A. From *in silico* target prediction to multi-target drug design: Current databases, methods and applications. *J. Proteomics* **2011**, *74* (12), 2554–2574.

[457] Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for molecular docking: A review. *Biophys. Rev.* **2017**, *9* (2), 91–102.

[458] Taylor, W. R. A "periodic table" for protein structures. *Nature* **2002**, *416* (6881), 657–660.

APPENDIX 1: Binding Site Prediction

Table A1.1. Overview of developed binding site prediction methods. The method name and the availability is given together with the corresponding URL whenever applicable. The basic approach for pocket detection (not ranking) is provided (G for geometry-based, E for energy-based, K for knowledge-based, and C for approaches which combine the results of more than one method) together with the year of publication. Methods analysed in this study are highlighted in grey (in light grey if a reimplementation was applied).

Method	Availability	URL	Approach	Year
Method of Kuntz *et al.*[1]	not available	n/a	G	1982
GRID[2]	not available	n/a	G	1985
Method of David[3]	not available	n/a	G	1988
Method of Voorintholt *et al.*[4]	not available	n/a	G	1989
FindSurf[5]	not available	n/a	G	1989
Cavity Search[6]	not available	n/a	G	1990
POCKET[7]	not available	n/a	G	1992
Method of Delaney[8]	not available	n/a	G	1992
Method of Del Carpio *et al.*[9]	not available	n/a	G	1993
ProACT[10]	standalone	http://people.cryst.bbk.ac.uk/~ubcg66a/proact2_summary.html	G/E	1994
VOIDOO[11]	standalone	http://xray.bmc.uu.se/usf/	G	1994
SURFNET[12]	standalone	http://www.ebi.ac.uk/thornton-srv/software/SURFNET/	G	1995
Sequence Space[13]	not available	n/a	K	1995
Method of Masuya & Doi[14]	not available	n/a	G	1995
APROPOS[15]	not available	http://www.csb.yale.edu/userguides/datamanip/apropos/manual.html	G	1996
MSMS[16]	standalone	http://mgltools.scripps.edu/downloads#msms	G	1996
Evolutionary Trace Server[17]	web server	http://mordred.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html	K	1996
Method of Ruppert *et al.*[18]	standalone	http://www.biopharmics.com/about.html	E	1997
LIGSITE[19]	not available	n/a	G	1997
Method of Exner *et al.*[20]	not available	n/a	G	1998
CAST[21]	not available	n/a	G	1998
vdW-FFT[22]	standalone	http://anusf.anu.edu.au/~aab900/dock.html	E	1999
PASS[23]	standalone	http://www.ccl.net/cca/software/UNIX/pass/overview.shtml	G	2000
Surface Patches[24]	not available	n/a	G	2000
Method of Elcock[25]	not available	n/a	E	2001
Method of Landgraf *et al.*[26]	not available	n/a	K	2001
Method of Aloy *et al.*[27]	not available	n/a	K	2001
MC method[28]	not available	n/a	G	2002
Rate4Site[29]	standalone	http://www.tau.ac.il/~itaymay/cp/rate4site.html	K	2002
CASTp[30,31]	web server	http://sts.bioe.uic.edu/castp/index.php	K	2003
ConSurf[32]	web server	http://consurf.tau.ac.il/2016/	K	2003
PINTS[33]	web server	http://www.russelllab.org/cgi-bin/tools/pints.pl	K	2003
SUMO[34]	web server	http://sumo-pbil.ibcp.fr/cgi-bin/sumo-help	K	2003

Table A1.1 *(continued)*. Overview of developed binding site prediction methods. The method name and the availability is given together with the corresponding URL whenever applicable. The basic approach for pocket detection (not ranking) is provided (G for geometry-based, E for energy-based, K for knowledge-based, and C for approaches which combine the results of more than one method) together with the year of publication. Methods analysed in this study are highlighted in grey (in light grey if a reimplementation was applied).

Method	Availability	URL	Approach	Year
MTreedet[35]	standalone	http://csbg.cnb.csic.es/pazos/Xdet/	K	2003
WebFEATURE[36]	not available	n/a	K	2003
LigandFit[37]	BIOVIA (standalone)	http://3dsbiovia.com/	G	2003
pvSOAR[38]	not available	n/a	K	2004
CRESCENDO[39]	web server	http://www.bioinf.manchester.ac.uk/crescendo/	K	2004
OMD[40]	not available	n/a	G/E	2004
DrugSite[41]	not available	n/a	E	2004
Method of Panchenko *et al.*[42]	not available	n/a	K	2004
CFG (Conserved Functional Group)[43]	not available	n/a	K	2004
Method of Bate & Warwicker[44]	not available	n/a	E	2004
Method of Amitai *et al.*[45]	not available	n/a	E	2004
ProFunc[46]	web server	http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/	G/K	2005
Q-SiteFinder[47]	not available	n/a	E	2005
PocketFinder (ICM)[48]	molsoft (standalone)	http://www.molsoft.com/	E	2005
PatchFinder[49]	web server	http://patchfinder.tau.ac.il/	K	2005
MEDock[50]	web server	http://bioinfo.mc.ntu.edu.tw/medock/	E	2005
EnSite[51]	not available	n/a	G	2005
PINUP[52]	standalone	http://sysbio.unl.edu/services/PINUP/	G/E	2006
LigSite^{CSC}[53]	web server	http://projects.biotec.tu-dresden.de/cgi-bin/index.php	G/K	2006
SCREEN2[54]	web server	https://bhapp.c2b2.columbia.edu/screen2/cgi-bin/screen2.cgi	G	2006
Protemot	web server	http://protemot.csbb.ntu.edu.tw/	K	2006
MCdet[55]	standalone	http://csbg.cnb.csic.es/pazos/Xdet/	K	2006
Xdet[55]	standalone	http://csbg.cnb.csic.es/pazos/Xdet/	K	2006
LIGSITE^{CS}[53]	standalone	http://projects.biotec.tu-dresden.de/pocket/download.html	G	2006
Travel Depth[56]	standalone	https://github.com/ryancoleman/traveldistance	G	2006
DPA[57]	not available	n/a	G/E	2006
CAVER[58]	standalone	https://www.caver.cz/index.php	G	2006
eF-seek[59]	web server	http://ef-site.hgc.jp/eF-seek/top.do	G/E	2007
firestar[60]	web server	http://firedb.bioinfo.cnio.es/Php/FireStar.php	K	2007
THEMATICS[61]	web server	http://www.pool.neu.edu/wPOOL/	E	2007
Pockets/Tunnel Extraction (TexMol)[62]	standalone	https://cvcweb.ices.utexas.edu/cvcwp/software/	G	2007
PAR-3D[63]	not available	n/a	K	2007
HotPatch[64]	not available	n/a	K	2007
SiteMap[65]	Schrödinger (standalone)	https://www.schrodinger.com/	E	2007

Table A1.1 *(continued)*. Overview of developed binding site prediction methods. The method name and the availability is given together with the corresponding URL whenever applicable. The basic approach for pocket detection (not ranking) is provided (G for geometry-based, E for energy-based, K for knowledge-based, and C for approaches which combine the results of more than one method) together with the year of publication. Methods analysed in this study are highlighted in grey (in light grey if a reimplementation was applied).

Method	Availability	URL	Approach	Year
Fuzzy Oil Drop[66]	web server	http://bioinformatics.cm-uj.krakow.pl/activesite/	E	2007
siteFiNDER|3D[67]	not available	n/a	K	2007
EPOS^{BP}[68]	standalone	http://gepard.bioinformatik.uni-saarland.de/software/epos-bp	G	2007
PHECOM[69]	standalone	http://strcomp.protein.osaka-u.ac.jp/ghecom/download_src.html	G	2007
PocketPicker[70]	standalone	http://gecco.org.chemie.uni-frankfurt.de/pocketpicker/download.html	G	2007
Binding Response[71]	standalone	http://cadd.umaryland.edu/CADD_resources.shtml	E	2007
Method of Xie & Bourne[72]	not available	n/a	G	2007
MOLE[73]	standalone	https://webchem.ncbr.muni.cz/Platform/App/Mole	G	2007
PLB[74]	not available	n/a	K	2007
Method of Tseng & Liang[75]	not available	n/a	K	2007
VisGrid[76]	standalone	http://kiharalab.org/VisGrid/	G	2008
FINDSITE[77]	web server / standalone	http://cssb.biology.gatech.edu/findsite	K	2008
PocketDepth[78]	web server	http://proline.physics.iisc.ernet.in/pocketdepth/	G	2008
INTREPID[79]	standalone	http://phylogenomics.berkeley.edu/software	K	2008
Method of Morita *et al.*[80]	not available	n/a	E	2008
multi-RELIEF[81]	web server	http://www.ibi.vu.nl/programs/multirelief/	K	2008
Fast DPA[82]	not available	n/a	G/E	2008
N-Func[83]	web server	http://patchfinder.tau.ac.il/N-Func2/	G/K	2008
Method of Bock *et al.*[84]	not available	n/a	G/K	2008
SitePredict[85]	not available	n/a	G/K	2008
FRcons[86]	not available	n/a	K	2008
HOLLOW[87]	standalone	http://hollow.sourceforge.net	G	2008
Method of Kim *et al.*[88]	not available	n/a	G	2008
metapocket[89]	web server	http://projects.biotec.tu-dresden.de/metapocket/	C	2009
FTMap[90]	web server	http://ftmap.bu.edu/login.php	E	2009
EDTSurf[91]	standalone	https://zhanglab.ccmb.med.umich.edu/EDTSurf/	G	2009
fpocket[92]	standalone	https://sourceforge.net/projects/fpocket/files/fpocket-1.0/fpocket-src-1.0/	G	2009
Voronoia[93]	standalone	http://bioinformatics.charite.de/voronoia/index.php?site=download	G	2009
SiteHound[94]	standalone	http://scbx.mssm.edu/sitehound/sitehound-download/download.html	E	2009
ConCavity[95]	standalone	http://compbio.cs.princeton.edu/concavity/	C	2009

Table A1.1 *(continued)*. Overview of developed binding site prediction methods. The method name and the availability is given together with the corresponding URL whenever applicable. The basic approach for pocket detection (not ranking) is provided (G for geometry-based, E for energy-based, K for knowledge-based, and C for approaches which combine the results of more than one method) together with the year of publication. Methods analysed in this study are highlighted in grey (in light grey if a reimplementation was applied).

Method	Availability	URL	Approach	Year
Mark-Us[96]	web server	http://honig.c2b2.columbia.edu/markus	K	2009
Binding Balls[97]	not available	n/a	G	2009
Method of Liu *et al.*[98]	not available	n/a	K	2009
WSsas[99]	web service	https://www.ebi.ac.uk/thornton-srv/databases/WSsas/	K	2009
POOL[100]	not available	n/a	G/E/K	2009
SplitPocket[101]	web server	http://pocket.med.wayne.edu/patch/	G/K	2009
fPOP[102]	web server	http://pocket.med.wayne.edu/fpop/	K	2010
3DLigandSite[103]	web server	http://www.sbg.bio.ic.ac.uk/3dligandsite/	K	2010
ProBiS[104]	web server	http://probis.cmm.ki.si/	K	2010
STP[105]	web server	http://opus.bch.ed.ac.uk/stp/	K	2010
POCASA[106]	web server	http://altair.sci.hokudai.ac.jp/g6/service/pocasa/	G	2010
DISCERN[107]	standalone	http://phylogenomics.berkeley.edu/software	G/K	2010
VICE[108]	not available	n/a	G	2010
DoGSite[109]	standalone	https://www.biosolveit.de/DoGSiteScorer/	G	2010
ghecom[110]	standalone	http://strcomp.protein.osaka-u.ac.jp/ghecom/download_src.html	G	2010
McVol[111]	standalone	http://www.bisb.uni-bayreuth.de/index.php?page=data/mcvol/mcvol	G	2010
fpocket2[112]	standalone	https://sourceforge.net/projects/fpocket/files/latest/download	G	2010
SCREEN + PCA[113]	not available	n/a	K	2010
IBIS[114]	web server	https://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi?search=2OCJA&tab=2	K	2010
CAVE[115]	standalone	http://cpc.cs.qub.ac.uk/summaries/AEHC_v1_0.html	G	2010
BUMBLE[116]	web server	http://bumble.hgc.jp/	E	2010
3V[117]	standalone/ web server	http://3vee.molmovdb.org/sourcecode.php	G	2010
Method of Phillips *et al.*[118]	not available	n/a	G	2010
TexMol[119]	standalone	https://cvcweb.ices.utexas.edu/cvcwp/software/texmol/	G	2011
AADS[120]	web server	http://www.scfbio-iitd.res.in/dock/ActiveSite_new.jsp	G	2011
SPACER[121]	web server	http://allostery.bii.a-star.edu.sg/	G	2011
FunFOLD[122]	web server / standalone	http://www.reading.ac.uk/bioinf/FunFOLD/FunFOLD_form.html	K	2011
ResidueDepth (DEPTH)[123]	web server / standalone	http://cospi.iiserpune.ac.in/depth/htdocs/index.html	G	2011
miniMINER[124]	not available	n/a	K	2011

Table A1.1 *(continued)*. Overview of developed binding site prediction methods. The method name and the availability is given together with the corresponding URL whenever applicable. The basic approach for pocket detection (not ranking) is provided (G for geometry-based, E for energy-based, K for knowledge-based, and C for approaches which combine the results of more than one method) together with the year of publication. Methods analysed in this study are highlighted in grey (in light grey if a reimplementation was applied).

Method	Availability	URL	Approach	Year
fconv[125]	standalone	http://pc1664.pharmazie.uni-marburg.de/download/	G	2011
MSPocket[126]	standalone	http://projects.biotec.tu-dresden.de/MSPocket/	G	2011
PocketAnalyzer^{PCA}[127]	standalone	https://sourceforge.net/projects/papca/	G	2011
Molsite[128]	standalone	http://presto.protein.osaka-u.ac.jp/myPresto4/index.php?lang=en	E	2011
Blind Docking[129]	not available	n/a	E	2011
Union Ball[130]	upon request	n/a	G	2011
metapocket2[131]	web server	http://projects.biotec.tu-dresden.de/metapocket/	C	2011
AutoMatch[132]	not available	n/a	E	2011
Method of Giard *et al.*[133]	not available	n/a	G	2011
VoroProt[134]	standalone	http://www.bti.vu.lt/en/departments/department-of-bioinformatics/software/voroprot	G	2011
dxTuber[135]	standalone	https://github.com/willhelm-mueller/dxTuber	G	2011
MDpocket[136]	standalone	http://fpocket.sourceforge.net/	G	2011
Method of Krone *et al.*[137]	not available	n/a	G	2011
Method of Lindow *et al.*[138]	not available	n/a	G	2011
FTSite[139]	web server	http://ftsite.bu.edu/	E	2012
dPredGB[140]	standalone	http://www.t38.ph.tum.de/index.php?id=88	C	2012
Frustratometer[141]	web server	http://lfp.qb.fcen.uba.ar/embnet/charge.php?app=Frustratometer&chains=NO	E	2012
RaptorXBinding[142]	web server	http://raptorx.uchicago.edu/BindingSite/	K	2012
FREPS[143]	not available	n/a	K	2012
LISE[144]	standalone	https://sourceforge.net/projects/msdock/files/	K	2012
COFACTOR[145]	web server / standalone	https://zhanglab.ccmb.med.umich.edu/COFACTOR/	K	2012
PDBinder[146]	standalone	n/a	K	2012
ProACT2 (Olsson & Williams, unpublished)	standalone	http://people.cryst.bbk.ac.uk/~ubcg66a/proact2_summary.html	G/E	2012
POOL server[147]	not available	n/a	C	2012
SPRITE[148]	web server	http://mfrlab.org/grafss/sprite/	K	2012
GIRAF[149]	standalone	n/a	K	2012
PSC[150]	not available	n/a	K	2012
RobustVoids[151]	web server	http://vgl.serc.iisc.ernet.in/robustCavities/index.html	G	2013
DrosteP[152]	standalone (Windows)	http://www1.na.icb.cnr.it/project/drosteppy/	K	2013

Table A1.1 *(continued)*. Overview of developed binding site prediction methods. The method name and the availability is given together with the corresponding URL whenever applicable. The basic approach for pocket detection (not ranking) is provided (G for geometry-based, E for energy-based, K for knowledge-based, and C for approaches which combine the results of more than one method) together with the year of publication. Methods analysed in this study are highlighted in grey (in light grey if a reimplementation was applied).

Method	Availability	URL	Approach	Year
eFindSite[153]	web server / standalone	http://brylinski.cct.lsu.edu/efindsite	K	2013
FTFlex[154]	web server	https://ftflex.bu.edu/	E	2013
Adepth[155]	web server	http://biodev.cea.fr/adepth/	G	2013
MCPath[156]	web server	http://safir.prc.boun.edu.tr/clbet_server/	E	2013
FunFOLD2[157]	web server / standalone	http://www.reading.ac.uk/bioinf/FunFOLD/FunFOLD_form_2_0.html	K	2013
TargetS[158]	web server	http://www.csbio.sjtu.edu.cn:8080/TargetS/	K	2013
Method of Wong *et al.*[159]	not available	n/a	K	2013
CAVITATOR[160]	standalone	http://cssb.biology.gatech.edu/Cavitator	G	2013
S-SITE[161]	web server / standalone	https://zhanglab.ccmb.med.umich.edu/COACH/	K	2013
TM-SITE[161]	web server / standalone	https://zhanglab.ccmb.med.umich.edu/COACH/	K	2013
COACH[161]	web server / standalone	https://zhanglab.ccmb.med.umich.edu/COACH/	C	2013
PLB-SAVE[162]	web server	http://save.cs.ntou.edu.tw/index.php	G	2013
CAVITY[163]	not available	n/a	G	2013
CMCS-BSP[164]	not available	n/a	C	2013
VolArea[165]	VMD plugin	http://www.fc.up.pt/PortoBioComp/Software/Volarea/Home.html	G	2013
MOLE2[166]	standalone	https://webchem.ncbr.muni.cz/Platform/App/Mole	G	2013
Method of Parulek *et al.*[167]	not available	n/a	G	2013
Method of Krone *et al.* 2[168]	not available	n/a	G	2013
Method of Lindow *et al.* 2[169]	not available	n/a	G	2013
NanoShaper[170]	standalone	https://www.electrostaticszone.eu/downloads	G	2013
TRAPP[171]	web server	https://trapp.h-its.org/	G	2013
Method of Wang *et al.*[172]	not available	n/a	K	2013
pocketZebra[173]	web server	https://biokinet.belozersky.msu.ru/pocketzebra	K	2014
GalaxySite[174]	web server	http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=SITE	K	2014
Method of Saberi Fathi & Tuszynski[175]	not available	n/a	C	2014
CCCPP[176]	standalone	http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html	G	2014
BetaVoid[177]	standalone	http://voronoi.hanyang.ac.kr/software.htm#BetaVoid	G	2014
KVFinder[178]	standalone	https://sourceforge.net/projects/kvfinder/	G	2014
CAVER Analyst[179]	standalone	https://www.caver.cz/index.php	G	2014
LES[180]	not available	n/a	G	2014

Table A1.1 *(continued)*. Overview of developed binding site prediction methods. The method name and the availability is given together with the corresponding URL whenever applicable. The basic approach for pocket detection (not ranking) is provided (G for geometry-based, E for energy-based, K for knowledge-based, and C for approaches which combine the results of more than one method) together with the year of publication. Methods analysed in this study are highlighted in grey (in light grey if a reimplementation was applied).

Method	Availability	URL	Approach	Year
trj_cavity[181]	standalone	https://sourceforge.net/projects/trjcavity/files/latest/download	G	2014
Method of Krone *et al.* 3[182]	standalone	https://svn.vis.uni-stuttgart.de/trac/megamol	G	2014
Method of Mirceva & Kulakov[183]	not available	n/a	K	2014
BetaCavityWeb[184]	web server	http://voronoi.hanyang.ac.kr/betacavityweb/	G	2015
LIBRA[185]	standalone (GUI)	http://www.computationalbiology.it/software.html	K	2015
OSML[186]	not available	n/a	K	2015
GetCleft[187]	standalone	http://biophys.umontreal.ca/nrg/NRG/Resources.html	G	2015
PrinCCes[188]	standalone	https://github.com/CzirjakGabor/PrinCCes	G	2015
P2RANK[189]	standalone	https://github.com/rdk/p2rank	K	2015
VISM-CFA[190]	not available	n/a	E	2015
CAVE-CL[191]	standalone	http://cpc.cs.qub.ac.uk/summaries/AEHC_v2_0.html	G	2015
mkgrid[192]	not available	n/a	G	2015
SOM-BSfinder[193]	standalone	https://github.com/bougui505/SOM/tree/SOM3D_dev	C	2015
LigandDSES[194]	not available	www2.ahu.edu.cn/pchen/web/LigandDSES.htm	K	2016
bSiteFinder[195]	web server	http://binfo.shmtu.edu.cn/bsitefinder/index.php?s=/Index/index	K	2016
Frustratometer2[196]	web server	http://frustratometer.qb.fcen.uba.ar	E	2016
LigandRF[197]	not available	n/a	K	2016
AutoSite[198]	standalone	http://adfr.scripps.edu/AutoDockFR/downloads.html	E	2016
CryptoSite[199]	web server / standalone	https://modbase.compbio.ucsf.edu/cryptosite/	K	2016
FunFOLD3[200]	standalone	https://www.reading.ac.uk/bioinf/FunFOLD/	K	2016
ISMBLab-LIG[201]	web server	http://ismblab.genomics.sinica.edu.tw/	K	2016
Method of Tsujikawa *et al.*[202]	not available	n/a	E/K	2016
EXPOSITE[203]	standalone	http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005293	G	2016
MF-PLB[204]	not available	n/a	K	2016
UOBPRM[205]	not available	n/a	K	2016
Method of Kim *et al.* 2[206]	not available	n/a	G	2016
Method of Jurcik *et al.*[207]	not available	n/a	G	2016
DeepSite[208]	web server	http://www.playmolecule.org/deepsite/	K	2017
TRAPP[209]	web server	http://trapp.h-its.org/	C	2017

Table A1.1 *(continued)*. Overview of developed binding site prediction methods. The method name and the availability is given together with the corresponding URL whenever applicable. The basic approach for pocket detection (not ranking) is provided (G for geometry-based, E for energy-based, K for knowledge-based, and C for approaches which combine the results of more than one method) together with the year of publication. Methods analysed in this study are highlighted in grey (in light grey if a reimplementation was applied).

Method	Availability	URL	Approach	Year
G-LoSA[210]	standalone	https://compbio.lehigh.edu/GLoSA/toolkit.html	K	2017
COFACTOR2[211]	web server / standalone	http://zhanglab.ccmb.med.umich.edu/COFACTOR/	K	2017
R-subspace[212]	not available	n/a	C	2017
GaussianFinder[213]	standalone	https://sourceforge.net/projects/gaussianfinder/	G	2017
CriticalFinder[214]	not available	n/a	G	2017

Table A1.2. The extended PDBbind core set. The characteristics of all structures of the extended PDBbind core set are shown together with the affinity/activity of the ligands. The logKa values were extracted from the PDBbind database[215].

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity (logKa)
1	3bub	A	apo	Q24451	3.2.1.114				
1	3ejr	A	holo	Q24451	3.2.1.114	HN4	1049	A	8.57
1	3d4z	A	holo	Q24451	3.2.1.114	GIM	1048	A	4.89
1	1ps3	A	holo	Q24451	3.2.1.114	KIF	1501	A	2.28
2	2qly	A	apo	O43451	3.2.1.20				
2	3l4u	A	holo	O43451	3.2.1.20	DSK	4001	A	7.52
2	3l4w	A	holo	O43451	3.2.1.20	MIG	1001	A	6.00
2	2qmj	A	holo	O43451	3.2.1.20	ACR	1001	A	4.21
3	2gpn	A	apo	P00489	2.4.1.2				
3	3g2n	A	holo	P00489	2.4.1.1	OAK	998	A	4.09
3	4mi9	A	holo	P00489	2.4.1.1	26W	901	A	3.42
3	3l7b	A	holo	P00489	2.4.1.1	DKZ	843	A	2.40
4	2cho	A	apo	Q89ZI2	3.2.1.170				
4	2vvn	A	holo	Q89ZI2	3.2.1.169	NHT	1718	A	7.30
4	2wca	A	holo	Q89ZI2	3.2.1.169	NP6	900	A	5.60
4	2w66	A	holo	Q89ZI2	3.2.1.169	HQ6	1716	A	4.05
5	2x8y	A	apo	Q10714	3.4.15.2				
5	2x8z	A	holo	Q10714	3.4.15.1	X8Z	1615	A	7.96
5	2xhm	A	holo	Q10714	3.4.15.1	K26	1615	A	6.80
5	2x97	A	holo	Q10714	3.4.15.1	RX4	1615	A	5.66
6	2cbi	A	apo	Q0TR53	3.2.1.170				
6	2j62	A	holo	Q0TR53	3.2.1.169	GSZ	1626	A	11.34
6	2cbj	A	holo	Q0TR53	3.2.1.169	OAN	1625	A	8.27
6	2x0y	A	holo	Q0TR53	3.2.1.169	X0T	1625	A	4.60
7	2iul	A	apo	P12821	3.4.15.2				
7	3l3n	A	holo	P12821	3.4.15.1	LSW	1	A	8.18
7	3bkl	A	holo	P12821	3.4.15.1	KAW	800	A	6.17
7	3bkk	A	holo	P12821	3.4.15.1	KAF	704	A	6.08
8	1ea5	A	apo	P04058	3.1.1.8				
8	1e66	A	holo	P04058	3.1.1.7	HUX	803	A	9.89
8	1h23	A	holo	P04058	3.1.1.7	E12	1536	A	8.35
8	1gpk	A	holo	P04058	3.1.1.7	HUP	1540	A	5.37

Table A1.2 *(continued)*. The extended PDBbind core set. The characteristics of all structures of the extended PDBbind core set are shown together with the affinity/activity of the ligands. The logKa values were extracted from the PDBbind database[215].

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity (logKa)
9	2zku	A	apo	P26663	2.7.7.49				
9	2gir	A	holo	P26663	2.7.7.48	NN3	1001	A	n/a
9	2d3u	A	holo	Q99AU2	2.7.7.48	CCT	1001	A	6.92
9	2d3z	A	holo	Q99AU2	2.7.7.48	FIH	2001	A	6.64
10	2a65	A	apo	O67854	n/a				
10	2q72	A	holo	O67854	n/a	IXX	801	A	n/a
10	2qb4	A	holo	O67854	n/a	DSM	801	A	n/a
10	2qei	A	holo	O67854	n/a	CXX	801	A	n/a
11	1hny	A	apo	P04746	3.2.1.1				
11	1xd0	A	holo	P04746	3.2.1.1	ARE	501	A	7.12
11	1u33	A	holo	P04746	3.2.1.1	LM2	500	A	4.60
11	4gqr	A	holo	P04746	3.2.1.1	MYC	504	A	3.96
12	1od0	A	apo	Q08638	3.2.1.22				
12	2cet	A	holo	Q08638	3.2.1.21	PGI	1447	A	8.02
12	2j78	A	holo	Q08638	3.2.1.21	GOX	1451	A	6.42
12	2wbg	A	holo	Q08638	3.2.1.21	LGS	1446	A	4.45
13	1hl8	A	apo	Q9WYE2	3.2.1.52				
13	2zx6	A	holo	Q9WYE2	3.2.1.51	ZX6	901	A	10.60
13	2zwz	A	holo	Q9WYE2	3.2.1.51	ZWZ	901	A	7.79
13	2zxd	A	holo	Q9WYE2	3.2.1.51	ZXD	901	A	5.22
14	3tpj	A	apo	P56817	3.4.23.47				
14	4gid	A	holo	P56817	3.4.23.46	0GH	501	A	10.77
14	4djv	A	holo	P56817	3.4.23.46	0KM	501	A	6.72
14	3udh	A	holo	P56817	3.4.23.46	091	501	A	2.85
15	2o15	A	apo	P9WPY5	2.5.1.19				
15	2o0b	A	holo	P9WPY5	2.5.1.19	S3P	702	A	n/a
15	2o0e	A	holo	P9WPY5	2.5.1.19	S3P, PEP	501,502	A, A	n/a
15	2o0x	A	holo	P9WPY5	2.5.1.19	SKP	501	A	n/a
16	7nn9	A	apo	P03472	3.2.1.18				
16	1f8c	A	holo	P03472	3.2.1.18	4AM	4	A	7.40
16	1f8b	A	holo	P03472	3.2.1.18	DAN	0	A	5.40
16	1f8d	A	holo	P03472	3.2.1.18	9AM	0	A	3.40
17	1pud	A	apo	P28720	2.4.2.29				
17	3ge7	A	holo	P28720	2.4.2.29	AFQ	390	A	8.70
17	1r5y	A	holo	P28720	2.4.2.29	DQU	900	A	6.46
17	1n2v	A	holo	P28720	2.4.2.29	BDI	900	A	4.08
18	1r39	A	apo	Q16539	2.7.11.24				
18	3e93	A	holo	Q16539	2.7.11.24	19B	361	A	8.85
18	3gcs	A	holo	Q16539	2.7.11.24	BAX	401	A	7.25
18	3huc	A	holo	Q16539	2.7.11.24	G97	362	A	5.99
19	2gfc	A	apo	P00517	2.7.11.11				
19	3zo4	A	holo	P00517	2.7.11.11	QWI	1352	A	n/a
19	1q8u	A	holo	P00517	2.7.11.11	H52	961	A	5.96
19	1q8t	A	holo	P00517	2.7.11.11	Y27	930	A	4.76
20	3fwq	A	apo	P68400	2.7.11.1				
20	3pe2	A	holo	P68400	2.7.11.1	E1B	338	A	9.76
20	2zjw	A	holo	P68400	2.7.11.1	REF	336	A	7.70
20	3owj	A	holo	P68400	2.7.11.1	1EL	332	A	6.07

Table A1.2 *(continued)*. The extended PDBbind core set. The characteristics of all structures of the extended PDBbind core set are shown together with the affinity/activity of the ligands. The logKa values were extracted from the PDBbind database[215].

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity (logKa)
21	3urj	A	apo	P11838	3.4.23.22				
21	1gvt	A	holo	P11838	3.4.23.22	2ZS	400	A	n/a
21	3pww	A	holo	P11838	3.4.23.22	ROC	1001	A	7.32
21	2v00	A	holo	P11838	3.4.23.22	V15	1336	A	3.66
22	2zav	A	apo	P05089	3.5.3.1				
22	3kv2	A	holo	P05089	3.5.3.1	NNH	901	A	7.32
22	3f80	A	holo	P05089	3.5.3.1	6HN	551	A	4.22
22	3mfv	A	holo	P05089	3.5.3.1	Z70	0	A	2.52
23	1oes	A	apo	P18031	3.1.3.48				
23	2qbp	A	holo	P18031	3.1.3.48	527	800	A	8.40
23	2qbr	A	holo	P18031	3.1.3.48	910	800	A	6.33
23	2hb1	A	holo	P18031	3.1.3.48	512	800	A	3.80
24	3t2h	E	apo	P00800	3.4.24.27				
24	4tmn	E	holo	P00800	3.4.24.27	0PK	317	E	10.17
24	1os0	A	holo	P00800	3.4.24.27	0PQ	605	A	6.03
24	3fcq	A	holo	P00800	3.4.24.27	M3S	600	A	2.77
25	3pxr	A	apo	P24941	2.7.11.22				
25	2r3q	A	holo	P24941	2.7.11.22	5SC	501	A	n/a
25	2fvd	A	holo	P24941	2.7.11.22	LIA	299	A	8.52
25	2xnb	A	holo	P24941	2.7.11.22	Y8L	1299	A	6.83
26	1qhz	A	apo	O85465	3.2.1.4				
26	1w3l	A	holo	O85465	3.2.1.4	BGC, BGC, OXZ	1307, 1308, 1309	A, A, A	6.28
26	1w3k	A	holo	O85465	3.2.1.4	BGC, OXZ	1304, 1305	A, A	4.30
26	1qi0	A	holo	O85465	3.2.1.4	CBI	501	A	2.35
27	3cov	A	apo	P9WIL5	6.3.2.1				
27	3coy	A	holo	P9WIL5	6.3.2.1	53H	301	A	6.02
27	3iub	A	holo	P9WIL5	6.3.2.1	FG2	302	A	4.54
27	3imc	A	holo	P9WIL5	6.3.2.1	BZ3	702	A	2.96
28	1rtq	A	apo	Q01693	3.4.11.10				
28	3vh9	A	holo	Q01693	3.4.11.10	HQY	303	A	6.24
28	3b3w	A	holo	Q01693	3.4.11.10	LEU	297	A	4.19
28	3b3s	A	holo	Q01693	3.4.11.10	LEU	304	A	2.55
29	3kfa	A	apo	P00520	2.7.10.2				
29	1opj	A	holo	P00520	2.7.10.2	MYR	1	A	n/a
29	3k5v	A	holo	P00520	2.7.10.2	STJ	1	A	6.30
29	3mss	A	holo	P00520	2.7.10.2	MS7	516	A	4.66
30	1ia8	A	apo	O14757	2.7.11.1				
30	2yex	A	holo	O14757	2.7.11.1	YEX	1278	A	n/a
30	1nvq	A	holo	O14757	2.7.11.1	UCN	400	A	8.25
30	2brb	A	holo	O14757	2.7.11.1	PFQ	1277	A	4.86
31	2zco	A	apo	A9JQL9	2.5.1.96				
31	2zcq	A	holo	A9JQL9	2.5.1.96	B65	451	A	8.82
31	2zcr	A	holo	A9JQL9	2.5.1.96	B69	668	A	6.87
31	3acw	A	holo	A9JQL9	2.5.1.96	651	1001	A	4.76

Table A1.2 *(continued)*. The extended PDBbind core set. The characteristics of all structures of the extended PDBbind core set are shown together with the affinity/activity of the ligands. The logKa values were extracted from the PDBbind database[215].

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity (logKa)
32	2uuf	B	apo	P00734	3.4.21.5				
32	3utu	H	holo	P00734	3.4.21.5	1TS	901	A	10.92
32	1oyt	H	holo	P00734	3.4.21.5	FSN	501	H	7.24
32	1bcu	H	holo	P00734	3.4.21.5	PRL	280	H	3.28
33	2zk0	A	apo	P37231	n/a				
33	3v9t	A	holo	P37231	n/a	17L	501	A	n/a
33	3b1m	A	holo	P37231	n/a	KRC	1	A	8.48
33	3u9q	A	holo	P37231	n/a	DKA	1	A	4.38
34	2bfx	A	apo	Q6DE08	2.7.11.1				
34	2vgo	A	holo	Q6DE08	2.7.11.1	AD5	1357	A	n/a
34	2vgp	A	holo	Q6DE08	2.7.11.1	AD6	1357	A	n/a
34	4b8m	A	holo	Q6DE08	2.7.11.1	VX6	1359	A	n/a
35	3oy5	U	apo	P00749	3.4.21.73				
35	1sqa	A	holo	P00749	3.4.21.73	UI1	1001	A	9.21
35	1o5b	B	holo	P00749	3.4.21.73	ESI	251	B	5.77
35	3kgp	A	holo	P00749	3.4.21.73	4AZ	4069	A	2.57
36	1ca2	A	apo	P00918	4.2.1.1				
36	3dd0	A	holo	P00918	4.2.1.1	EZL	301	A	9.00
36	2weg	A	holo	P00918	4.2.1.1	FBV, GOL	1263, 1264	A, A	6.50
36	3kwa	A	holo	P00918	4.2.1.1	SPM	300	A	4.08
37	3t0h	A	apo	P07900	n/a				
37	2yki	A	holo	P07900	n/a	YKI	1224	A	9.46
37	1yc1	A	holo	P07900	n/a	4BC	401	A	6.17
37	2xdl	A	holo	P07900	n/a	2DL	1225	A	3.10
38	1mxv	A	apo	P19491	n/a				
38	4g8m	A	holo	P19491	n/a	G8M	301	A	7.89
38	3bfu	A	holo	P19491	n/a	R2P	264	A	6.27
38	1p1q	A	holo	P19491	n/a	AMQ	602	A	4.89
39	2p74	A	apo	Q9L5C8	3.5.2.6				
39	4de1	A	holo	Q9L5C8	3.5.2.6	OJ6	301	A	5.96
39	4de2	A	holo	Q9L5C8	3.5.2.6	DN3	301	A	4.12
39	3g2z	A	holo	Q9L5C8	3.5.2.6	GZ2	1	A	2.36
40	4u2r	A	apo	P19491	n/a				
40	3o6g	A	holo	P19491	n/a	GLU	264	A	n/a
40	4u1z	A	holo	P19491	n/a	KAI	301	A	n/a
40	4u21	A	holo	P19491	n/a	FWD	401	A	n/a
41	1hcg	A	apo	P00742	3.4.21.6				
41	1mq6	A	holo	P00742	3.4.21.6	XLD	500	A	11.15
41	2xbv	A	holo	P00742	3.4.21.6	XBV	1245	A	8.43
41	2y5h	A	holo	P00742	3.4.21.6	Y5H	1244	A	5.79
42	1dv7	A	apo	O26232	4.1.1.23				
42	1lor	A	holo	O26232	4.1.1.23	BMP	2001	A	11.06
42	1lol	A	holo	O26232	4.1.1.23	XMP	2001	A	6.39
42	1loq	A	holo	O26232	4.1.1.23	U	2001	A	3.70

Table A1.2 *(continued)*. The extended PDBbind core set. The characteristics of all structures of the extended PDBbind core set are shown together with the affinity/activity of the ligands. The logKa values were extracted from the PDBbind database[215].

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity (logKa)
43	2g55	A	apo	P00760	3.4.21.4				
43	1o3f	A	holo	P00760	3.4.21.4	696	246	A	7.96
43	3gy4	A	holo	P00760	3.4.21.4	PBZ	1	A	5.10
43	1uto	A	holo	P00760	3.4.21.4	PEA	1246	A	2.27
44	3c0e	A	apo	P02829	n/a				
44	2vw5	A	holo	P02829	n/a	BC6	500	A	8.52
44	2iwx	A	holo	P02829	n/a	M1S	1215	A	6.68
44	2yge	A	holo	P02829	n/a	GDM	1215	A	5.06
45	2y7y	A	apo	Q8WSF8	n/a				
45	2x00	A	holo	Q8WSF8	n/a	GYN	301	A	11.33
45	2xys	A	holo	Q8WSF8	n/a	SY9	1206	A	7.42
45	2ymd	A	holo	Q8WSF8	n/a	SRO	1213	A	3.16
46	1q9k	B	apo	n/a	n/a				
46	1kel	H	holo	P01868	n/a	AAH	219	H	7.28
46	3bpc	B	holo	n/a	n/a	KDA	212	B	4.80
46	2r23	B	holo	n/a	n/a	K01, KDA	214, 215	B, B	3.72
47	4p7j	A	apo	P22734	2.1.1.6				
47	3nw9	A	holo	P22734	2.1.1.6	637	228	A	9.00
47	3oe5	A	holo	P22734	2.1.1.6	611	223	A	6.88
47	3ozt	A	holo	P22734	2.1.1.6	OZZ	223	A	4.13
48	4isv	A, B	apo	n/a	n/a				
48	1igj	A, B	holo	S52028, S38950	n/a	DGX	228	B	10.00
48	2pcp	C, D	holo	PC4203, n/a	n/a	1PC	212	C	8.70
48	1zea	H, L	holo	Q921A6, A2NHM3	n/a	DVA, DPR, GLY, DSN, DGN, DHI, DTY, DAS, DSN	1, 2, 3, 4, 5, 6, 7, 8, 9	A, A, A, A, A, A, A, A, A	5.22
49	3hkr	A	apo	P09211	2.5.1.18				
49	10gs	A	holo	P09211	2.5.1.18	VWW	210	A	6.40
49	2gss	A	holo	P09211	2.5.1.18	EAA	0	A	4.94
49	1lbk	A	holo	P09211	2.5.1.18	GSH	504	A	3.18
50	3rc6	A	apo	A8DG50	3.4.21.98				
50	3su3	A	holo	A8DG50	3.4.21.98	SU3	1	A	9.13
50	3su2	A	holo	A8DG50	3.4.21.98	TSV	100	A	7.35
50	3su5	A	holo	A8DG50	3.4.21.98	SU3	1	A	5.58
51	1h05	A	apo	P9WPX7	4.2.1.10				
51	2xb8	A	holo	P9WPX7	4.2.1.10	XNW	1144	A	7.59
51	3n86	A	holo	P9WPX7	4.2.1.10	RJP	147	A	5.64
51	3n7a	A	holo	P9WPX7	4.2.1.10	FA1	147	A	3.70

Table A1.2 *(continued)*. The extended PDBbind core set. The characteristics of all structures of the extended PDBbind core set are shown together with the affinity/activity of the ligands. The logKa values were extracted from the PDBbind database[215].

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity (logKa)
52	3l3u	A, B	apo	Q72498	2.7.7.0				
52	3zso	A, B	holo	Q76353	2.7.7.0	O2N	1219	A	5.12
52	3zsq	A, B	holo	Q76353	2.7.7.0	O4N	1215	A	4.96
52	3zsx	A, B	holo	Q76353	2.7.7.0	N44	1215	A	3.28
53	4lzu	A	apo	P02754	n/a				
53	3uex	A	holo	P02754	n/a	STE	163	A	6.92
53	3ueu	A	holo	P02754	n/a	DAO	2871	A	5.24
53	3nq3	A	holo	P02754	n/a	DKA	3053	A	3.78
54	2oxu	A	apo	P39900	3.4.24.65				
54	3f17	A	holo	P39900	3.4.24.65	HS4	0	A	8.63
54	3ehy	A	holo	P39900	3.4.24.65	TBL	1	A	5.85
54	3lka	A	holo	P39900	3.4.24.65	M4S, HAE	1, 269	A, A	2.82
55	3cfm	A	apo	P02766	n/a				
55	4dew	A	holo	P02766	n/a	LU2	201	A	7.00
55	4des	A	holo	P02766	n/a	57D	201	A	5.85
55	3cft	A	holo	P02766	n/a	5NS	1	A	4.19
56	1kf3	A	apo	P61823	3.1.27.5				
56	1u1b	A	holo	P61823	3.1.27.5	PAX	901	A	7.80
56	1w4o	A	holo	P61823	3.1.27.5	UA3	1125	A	5.22
56	3dxg	A	holo	P61823	3.1.27.5	U5P, U5P	125, 126	A, A	2.40
57	2h46	E	apo	P62993	n/a				
57	1jyq	A	holo	P62993	n/a	MAZ, PTR, PTM, ASN	201, 202, 203, 204	L, L, L, L	8.70
57	3s8o	A	holo	P62993	n/a	ACE, PTR, 02K, ASN, NH2	1, 2, 3, 4, 5	B, B, B, B, B	6.85
57	3ov1	A	holo	P62993	n/a	ACT, PTR, 1AC, ASN, NH2	1, 2, 3, 4, 5	B, B, B, B, B	5.20
58	2g69	A	apo	Q7SSI0	3.4.23.16				
58	4djr	A	holo	Q90K99	3.4.23.16	OKJ	101	A	11.52
58	3cyx	A	holo	P04587	3.4.23.16	ROC	201	A	8.00
58	1a30	A	holo	P04585	3.4.23.16	GLU, ASP, LEU	506, 507, 508	C, C, C	4.30
59	1dp0	A	apo	P00722	3.2.1.23				
59	3vd4	A	holo	P00722	3.2.1.23	IPT	2001	A	4.82
59	3muz	1	holo	P00722	3.2.1.23	IPT	2001	1	3.46
59	3i3b	A	holo	P00722	3.2.1.23	149	2001	A	2.23

Table A1.2 *(continued)*. The extended PDBbind core set. The characteristics of all structures of the extended PDBbind core set are shown together with the affinity/activity of the ligands. The logKa values were extracted from the PDBbind database[215].

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity (logKa)
60	4jho	A	apo	A2YPH1	3.2.1.25				
60	4jie	A	holo	A2YPH1	3.2.1.25	BMA	508	A	n/a
60	4re4	A	holo	A2YPH1	3.2.1.25	IFM	501	A	6.87
60	4re2	A	holo	A2YPH1	3.2.1.25	MVL	501	A	4.98
61	2bgr	A	apo	P27487	3.4.14.5				
61	3nox	A	holo	P27487	3.4.14.5	6A5	1	A	8.66
61	2ole	A	holo	P27487	3.4.14.5	KR2	1001	A	7.25
61	1n1m	A	holo	P27487	3.4.14.5	A3M	954	A	5.70
62	3kpj	A	apo	P11086	2.1.1.28				
62	2obf	A	holo	P11086	2.1.1.28	F83	2001	A	8.85
62	2g70	A	holo	P11086	2.1.1.28	HNT	5001	A	7.77
62	1hnn	A	holo	P11086	2.1.1.28	SKF	3001	A	6.24
63	2am9	A	apo	P10275	n/a				
63	2pip	L	holo	P10275	n/a	ICO	935	L	n/a
63	2pit	A	holo	P10275	n/a	4HY	933	A	n/a
63	2piv	A	holo	P10275	n/a	T3	933	A	n/a
64	1qia	A	apo	P08254	3.4.24.17				
64	1hfs	A	holo	P08254	3.4.24.17	L04	256	A	8.70
64	2d1o	A	holo	P08254	3.4.24.17	FA4	1001	A	7.70
64	1sln	A	holo	P08254	3.4.24.17	INH	256	A	6.64
65	1ous	A	apo	Q9HYN5	n/a				
65	2jdu	C	holo	Q9HYN5	n/a	MFU	1117	C	6.72
65	2jdm	A	holo	Q9HYN5	n/a	FUC	1118	A	5.40
65	2jdy	A	holo	Q9HYN5	n/a	MMA	1115	A	4.37

Table A1.3. RMSD values between ligand-bound and unbound structures for the extended PDBbind core set. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches. The average RMSD (Ø) of the apo structure to all holo structures is also given.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
1	3bub	A	Ø	0.18	1.36	0.26
1	3ejr	A	holo	0.20	1.35	0.19
1	3d4z	A	holo	0.18	1.35	0.22
1	1ps3	A	holo	0.18	1.39	0.36
2	2qly	A	Ø	0.24	2.60	0.19
2	3l4u	A	holo	0.29	2.50	0.15
2	3l4w	A	holo	0.22	1.57	0.17
2	2qmj	A	holo	0.23	3.74	0.25
3	2gpn	A	Ø	0.51	0.32	0.47
3	3g2n	A	holo	0.41	0.30	0.45
3	4mi9	A	holo	0.56	0.18	0.57
3	3l7b	A	holo	0.57	0.49	0.38

Table A1.3 *(continued)*. RMSD values between ligand-bound and unbound structures for the extended PDBbind core set. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches. The average RMSD (Ø) of the apo structure to all holo structures is also given.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
4	2cho	A	Ø	0.48	2.56	0.56
4	2vvn	A	holo	0.22	2.69	0.96
4	2wca	A	holo	0.95	2.69	0.61
4	2w66	A	holo	0.28	2.31	0.12
5	2x8y	A	Ø	0.11	0.11	0.23
5	2x8z	A	holo	0.11	0.08	0.24
5	2xhm	A	holo	0.11	0.13	0.20
5	2x97	A	holo	0.13	0.13	0.25
6	2cbi	A	Ø	0.81	2.80	0.53
6	2j62	A	holo	1.04	2.83	0.73
6	2cbj	A	holo	0.91	2.50	0.69
6	2x0y	A	holo	0.49	3.06	0.16
7	2iul	A	Ø	0.24	0.10	0.24
7	3l3n	A	holo	0.11	0.13	0.22
7	3bkl	A	holo	0.25	0.12	0.25
7	3bkk	A	holo	0.36	0.04	0.23
8	1ea5	A	Ø	0.17	2.73	0.20
8	1e66	A	holo	0.18	0.17	0.21
8	1h23	A	holo	0.19	4.08	0.20
8	1gpk	A	holo	0.15	3.94	0.19
9	2zku	A	Ø	0.50	0.72	0.42
9	2gir	A	holo	0.54	0.70	0.47
9	2d3u	A	holo	0.47	0.72	0.41
9	2d3z	A	holo	0.50	0.73	0.37
10	2a65	A	Ø	0.22	2.03	0.32
10	2q72	A	holo	0.19	2.14	0.36
10	2qb4	A	holo	0.19	2.13	0.34
10	2qei	A	holo	0.28	1.81	0.27
11	1hny	A	Ø	0.41	3.14	0.64
11	1xd0	A	holo	0.36	3.41	0.46
11	1u33	A	holo	0.46	2.22	0.70
11	4gqr	A	holo	0.43	3.80	0.75
12	1od0	A	Ø	0.21	1.90	0.23
12	2cet	A	holo	0.15	1.04	0.24
12	2j78	A	holo	0.25	1.50	0.31
12	2wbg	A	holo	0.25	3.17	0.13
13	1hl8	A	Ø	0.62	1.55	0.65
13	2zx6	A	holo	0.62	2.18	0.70
13	2zwz	A	holo	0.61	2.00	0.65
13	2zxd	A	holo	0.64	0.46	0.61
14	3tpj	A	Ø	0.87	1.01	0.66
14	4gid	A	holo	0.83	1.43	0.67
14	4djv	A	holo	0.87	0.71	0.65
14	3udh	A	holo	0.92	0.89	0.67
15	2o15	A	Ø	0.64	3.49	0.90
15	2o0b	A	holo	0.65	3.41	0.90
15	2o0e	A	holo	0.65	3.40	0.94
15	2o0x	A	holo	0.62	3.66	0.85

Table A1.3 *(continued)*. RMSD values between ligand-bound and unbound structures for the extended PDBbind core set. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches. The average RMSD (Ø) of the apo structure to all holo structures is also given.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
16	7nn9	A	Ø	0.17	2.63	0.31
16	1f8c	A	holo	0.17	3.03	0.28
16	1f8b	A	holo	0.17	1.81	0.33
16	1f8d	A	holo	0.19	3.04	0.31
17	1pud	A	Ø	0.39	0.95	0.44
17	3ge7	A	holo	0.31	0.51	0.38
17	1r5y	A	holo	0.44	0.49	0.61
17	1n2v	A	holo	0.43	1.85	0.32
18	1r39	A	Ø	1.37	1.85	0.62
18	3e93	A	holo	1.33	1.98	0.76
18	3gcs	A	holo	1.35	2.09	0.69
18	3huc	A	holo	1.45	1.47	0.43
19	2gfc	A	Ø	0.62	1.21	0.44
19	3zo4	A	holo	0.59	1.42	0.40
19	1q8u	A	holo	0.69	1.22	0.54
19	1q8t	A	holo	0.60	0.99	0.39
20	3fwq	A	Ø	1.57	2.06	0.48
20	3pe2	A	holo	1.50	1.62	0.40
20	2zjw	A	holo	1.68	1.57	0.36
20	3owj	A	holo	1.53	2.99	0.67
21	3urj	A	Ø	0.52	1.26	0.35
21	1gvt	A	holo	0.23	0.28	0.36
21	3pww	A	holo	0.64	0.40	0.34
21	2v00	A	holo	0.69	3.10	0.35
22	2zav	A	Ø	0.19	3.02	1.05
22	3kv2	A	holo	0.22	3.45	1.17
22	3f80	A	holo	0.19	2.91	1.08
22	3mfv	A	holo	0.17	2.71	0.89
23	1oes	A	Ø	1.25	2.57	0.37
23	2qbp	A	holo	1.25	2.56	0.21
23	2qbr	A	holo	1.25	2.56	0.19
23	2hb1	A	holo	1.26	2.59	0.70
24	3t2h	E	Ø	0.62	0.60	0.44
24	4tmn	E	holo	0.63	0.57	0.48
24	1os0	A	holo	0.58	0.61	0.39
24	3fcq	A	holo	0.65	0.62	0.45
25	3pxr	A	Ø	1.03	0.60	0.54
25	2r3q	A	holo	0.69	0.57	0.22
25	2fvd	A	holo	1.42	0.31	0.45
25	2xnb	A	holo	1.00	0.93	0.94
26	1qhz	A	Ø	0.49	0.30	0.31
26	1w3l	A	holo	0.70	0.18	0.39
26	1w3k	A	holo	0.70	0.32	0.31
26	1qi0	A	holo	0.09	0.39	0.22
27	3cov	A	Ø	0.26	0.19	0.74
27	3coy	A	holo	0.50	0.33	0.86
27	3iub	A	holo	0.15	0.11	0.56
27	3imc	A	holo	0.15	0.14	0.79

Table A1.3 *(continued)*. RMSD values between ligand-bound and unbound structures for the extended PDBbind core set. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches. The average RMSD (Ø) of the apo structure to all holo structures is also given.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
28	1rtq	A	Ø	0.21	1.11	0.72
28	3vh9	A	holo	0.21	0.53	0.78
28	3b3w	A	holo	0.20	0.38	0.79
28	3b3s	A	holo	0.22	2.43	0.58
29	3kfa	A	Ø	1.35	2.99	0.30
29	1opj	A	holo	1.65	2.85	0.43
29	3k5v	A	holo	1.36	3.06	0.23
29	3mss	A	holo	1.05	3.05	0.23
30	1ia8	A	Ø	0.75	0.80	0.29
30	2yex	A	holo	1.51	0.32	0.47
30	1nvq	A	holo	0.37	0.33	0.18
30	2brb	A	holo	0.38	1.74	0.21
31	2zco	A	Ø	0.49	2.04	0.96
31	2zcq	A	holo	0.98	0.44	1.25
31	2zcr	A	holo	0.24	3.15	0.78
31	3acw	A	holo	0.25	2.52	0.86
32	2uuf	B	Ø	0.20	0.20	0.26
32	3utu	H	holo	0.24	0.14	0.33
32	1oyt	H	holo	0.14	0.15	0.25
32	1bcu	H	holo	0.23	0.31	0.20
33	2zk0	A	Ø	0.96	2.59	0.58
33	3v9t	A	holo	1.02	2.52	0.54
33	3b1m	A	holo	0.92	2.65	0.56
33	3u9q	A	holo	0.95	2.60	0.63
34	2bfx	A	Ø	0.15	0.20	0.30
34	2vgo	A	holo	0.15	0.20	0.35
34	2vgp	A	holo	0.15	0.20	0.37
34	4b8m	A	holo	0.16	0.20	0.17
35	3oy5	U	Ø	0.42	0.25	0.36
35	1sqa	A	holo	0.50	0.32	0.36
35	1o5b	B	holo	0.53	0.27	0.43
35	3kgp	A	holo	0.23	0.17	0.30
36	1ca2	A	Ø	0.28	0.17	1.69
36	3dd0	A	holo	0.18	0.22	1.72
36	2weg	A	holo	0.25	0.10	1.55
36	3kwa	A	holo	0.41	0.19	1.81
37	3t0h	A	Ø	0.89	2.10	0.30
37	2yki	A	holo	0.98	2.23	0.24
37	1yc1	A	holo	1.37	0.17	0.53
37	2xdl	A	holo	0.33	3.91	0.14
38	1mxv	A	Ø	0.56	0.27	0.41
38	4g8m	A	holo	0.70	0.06	0.34
38	3bfu	A	holo	0.52	0.41	0.37
38	1p1q	A	holo	0.47	0.34	0.51
39	2p74	A	Ø	0.19	0.19	0.24
39	4de1	A	holo	0.17	0.10	0.24
39	4de2	A	holo	0.25	0.22	0.33
39	3g2z	A	holo	0.17	0.24	0.14

Table A1.3 *(continued)*. RMSD values between ligand-bound and unbound structures for the extended PDBbind core set. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches. The average RMSD (Ø) of the apo structure to all holo structures is also given.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
40	4u2r	A	Ø	2.27	2.61	0.35
40	3o6g	A	holo	2.42	2.67	0.30
40	4u1z	A	holo	2.05	3.17	0.51
40	4u21	A	holo	2.35	2.00	0.25
41	1hcg	A	Ø	0.60	1.71	0.49
41	1mq6	A	holo	0.62	1.69	0.44
41	2xbv	A	holo	0.63	1.73	0.68
41	2y5h	A	holo	0.56	1.71	0.34
42	1dv7	A	Ø	0.70	3.19	0.26
42	1lor	A	holo	0.70	3.06	0.37
42	1lol	A	holo	0.63	3.53	0.24
42	1loq	A	holo	0.79	2.99	0.19
43	2g55	A	Ø	0.24	0.18	0.28
43	1o3f	A	holo	0.32	0.23	0.36
43	3gy4	A	holo	0.25	0.12	0.35
43	1uto	A	holo	0.15	0.20	0.12
44	3c0e	A	Ø	0.74	0.36	0.33
44	2vw5	A	holo	0.89	0.32	0.37
44	2iwx	A	holo	0.61	0.39	0.21
44	2yge	A	holo	0.72	0.37	0.41
45	2y7y	A	Ø	0.64	1.09	0.58
45	2x00	A	holo	0.56	0.06	0.29
45	2xys	A	holo	0.57	2.76	0.58
45	2ymd	A	holo	0.79	0.45	0.86
46	1q9k	B	Ø	2.44	0.94	0.51
46	1kel	H	holo	5.04	0.88	0.23
46	3bpc	B	holo	1.15	0.95	0.65
46	2r23	B	holo	1.15	0.99	0.64
47	4p7j	A	Ø	1.72	2.75	0.74
47	3nw9	A	holo	1.93	3.09	0.75
47	3oe5	A	holo	1.61	2.60	0.73
47	3ozt	A	holo	1.62	2.57	0.74
48	4isv	A, B	Ø	3.99	2.94	0.94
48	1igj	A, B	holo	4.12	2.55	1.94
48	2pcp	C, D	holo	2.10	2.79	0.58
48	1zea	H, L	holo	5.75	3.49	0.32
49	3hkr	A	Ø	0.44	0.62	0.16
49	10gs	A	holo	0.31	0.17	0.18
49	2gss	A	holo	0.19	1.54	0.13
49	1lbk	A	holo	0.84	0.15	0.15
50	3rc6	A	Ø	0.59	0.35	0.46
50	3su3	A	holo	0.62	0.34	0.59
50	3su2	A	holo	0.34	0.35	0.24
50	3su5	A	holo	0.81	0.36	0.55
51	1h05	A	Ø	0.40	0.46	0.61
51	2xb8	A	holo	0.37	0.39	0.60
51	3n86	A	holo	0.40	0.26	0.34
51	3n7a	A	holo	0.45	0.74	0.90

Table A1.3 *(continued)*. RMSD values between ligand-bound and unbound structures for the extended PDBbind core set. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches. The average RMSD (Ø) of the apo structure to all holo structures is also given.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
52	3l3u	A, B	Ø	1.95	1.51	0.51
52	3zso	A, B	holo	1.96	1.49	0.58
52	3zsq	A, B	holo	1.95	1.49	0.47
52	3zsx	A, B	holo	1.94	1.55	0.49
53	4lzu	A	Ø	1.59	1.24	0.43
53	3uex	A	holo	1.64	1.65	0.62
53	3ueu	A	holo	1.46	0.90	0.31
53	3nq3	A	holo	1.69	1.18	0.35
54	2oxu	A	Ø	0.23	0.31	0.48
54	3f17	A	holo	0.27	0.23	0.75
54	3ehy	A	holo	0.20	0.37	0.41
54	3lka	A	holo	0.24	0.33	0.28
55	3cfm	A	Ø	0.27	0.13	0.63
55	4dew	A	holo	0.21	0.12	0.62
55	4des	A	holo	0.43	0.20	1.15
55	3cft	A	holo	0.18	0.08	0.13
56	1kf3	A	Ø	0.56	0.32	0.22
56	1u1b	A	holo	0.7	0.31	0.27
56	1w4o	A	holo	0.38	0.34	0.12
56	3dxg	A	holo	0.61	0.30	0.26
57	2h46	E	Ø	1.03	0.38	0.49
57	1jyq	A	holo	1.01	0.40	0.49
57	3s8o	A	holo	1.04	0.37	0.48
57	3ov1	A	holo	1.06	0.37	0.48
58	2g69	A	Ø	1.16	1.63	0.53
58	4djr	A	holo	1.14	1.82	0.68
58	3cyx	A	holo	1.16	1.80	0.27
58	1a30	A	holo	1.19	1.26	0.64
59	1dp0	A	Ø	0.37	0.97	1.00
59	3vd4	A	holo	0.26	0.28	1.04
59	3muz	1	holo	0.27	2.52	1.03
59	3i3b	A	holo	0.59	0.11	0.93
60	4jho	A	Ø	0.17	n/a	0.82
60	4jie	A	holo	0.24	n/a	0.93
60	4re4	A	holo	0.17	n/a	0.87
60	4re2	A	holo	0.10	n/a	0.67
61	2bgr	A	Ø	0.39	2.02	0.28
61	3nox	A	holo	0.36	1.83	0.28
61	2ole	A	holo	0.39	3.27	0.30
61	1n1m	A	holo	0.43	0.96	0.27
62	3kpj	A	Ø	0.34	2.10	0.31
62	2obf	A	holo	0.29	0.55	0.35
62	2g70	A	holo	0.34	2.76	0.22
62	1hnn	A	holo	0.39	3.00	0.35
63	2am9	A	Ø	0.54	2.23	0.47
63	2pip	L	holo	0.49	1.57	0.44
63	2pit	A	holo	0.48	2.50	0.47
63	2piv	A	holo	0.65	2.63	0.49

Table A1.3 *(continued)*. RMSD values between ligand-bound and unbound structures for the extended PDBbind core set. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches. The average RMSD (Ø) of the apo structure to all holo structures is also given.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
64	1qia	A	Ø	0.73	0.79	0.73
64	1hfs	A	holo	0.55	0.56	0.88
64	2d1o	A	holo	0.97	0.62	0.51
64	1sln	A	holo	0.68	1.20	0.78
65	1ous	A	Ø	0.51	0.65	1.36
65	2jdu	C	holo	0.50	0.59	1.29
65	2jdm	A	holo	0.49	0.66	1.43
65	2jdy	A	holo	0.55	0.69	1.35

Table A1.4. DoGSite descriptors for the extended PDBbind core set. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
1	3bub	apo	19.52	9.79	0.55	0.16	0.66	0.23	0.16	0.16
1	1ps3	holo	20.32	9.62	0.54	0.16	0.64	0.25	0.16	0.16
1	3d4z	holo	16.64	9.16	0.55	0.16	0.65	0.24	0.17	0.17
1	3ejr	holo	23.20	9.19	0.51	0.19	0.67	0.23	0.15	0.15
2	2qly	apo	11.20	9.74	0.60	0.07	0.70	0.19	0.13	0.13
2	2qmj	holo	14.40	11.08	0.56	0.08	0.72	0.19	0.10	0.10
2	3l4u	holo	12.48	10.94	0.60	0.07	0.68	0.20	0.14	0.14
2	3l4w	holo	18.88	11.69	0.54	0.10	0.70	0.17	0.14	0.14
3	2gpn	apo	33.76	35.98	0.60	0.02	0.65	0.19	0.19	0.19
3	3g2n	holo	32.96	38.12	0.60	0.03	0.62	0.21	0.20	0.20
3	3l7b	holo	32.64	37.05	0.58	0.03	0.62	0.21	0.20	0.20
3	4mi9	holo	52.16	28.44	0.53	0.05	0.67	0.17	0.19	0.19
4	2cho	apo	17.12	15.45	0.56	0.07	0.75	0.14	0.13	0.13
4	2vvn	holo	41.60	20.22	0.52	0.13	0.73	0.15	0.14	0.14
4	2w66	holo	20.00	14.17	0.54	0.08	0.74	0.16	0.13	0.13
4	2wca	holo	80.96	17.94	0.49	0.15	0.72	0.15	0.14	0.14
5	2x8y	apo	198.88	37.16	0.42	0.14	0.65	0.21	0.19	0.19
5	2x8z	holo	322.24	38.62	0.47	0.12	0.66	0.19	0.19	0.19
5	2x97	holo	262.40	42.97	0.43	0.14	0.66	0.21	0.17	0.17
5	2xhm	holo	186.40	35.05	0.46	0.10	0.67	0.21	0.16	0.16
6	2cbi	apo	49.92	16.87	0.52	0.09	0.71	0.16	0.15	0.15
6	2cbj	holo	21.76	15.46	0.53	0.08	0.72	0.17	0.14	0.14
6	2j62	holo	21.76	15.80	0.57	0.05	0.74	0.15	0.14	0.14
6	2x0y	holo	52.32	17.04	0.52	0.09	0.74	0.15	0.13	0.13
7	2iul	apo	555.36	53.17	0.41	0.17	0.69	0.19	0.17	0.17
7	3bkk	holo	157.44	28.39	0.39	0.22	0.67	0.19	0.19	0.19
7	3bkl	holo	295.04	47.12	0.41	0.14	0.68	0.19	0.17	0.17
7	3l3n	holo	240.16	29.62	0.43	0.16	0.67	0.20	0.17	0.17
8	1ea5	apo	79.84	20.29	0.39	0.09	0.77	0.14	0.12	0.12
8	1e66	holo	64.48	19.86	0.42	0.06	0.76	0.15	0.13	0.13
8	1gpk	holo	61.92	19.93	0.41	0.06	0.77	0.14	0.13	0.13
8	1h23	holo	69.44	21.44	0.40	0.08	0.75	0.14	0.14	0.14

Table A1.4 *(continued)*. DoGSite descriptors for the extended PDBbind core set. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
9	2zku	apo	129.60	19.28	0.37	0.26	0.64	0.15	0.23	0.23
9	2d3u	holo	83.20	14.32	0.37	0.25	0.66	0.16	0.21	0.21
9	2d3z	holo	89.92	13.98	0.38	0.25	0.68	0.14	0.18	0.18
9	2gir	holo	121.92	19.68	0.36	0.26	0.68	0.16	0.18	0.18
10	2a65	apo	33.76	14.36	0.45	0.07	0.78	0.12	0.13	0.13
10	2q72	holo	41.12	16.34	0.46	0.08	0.75	0.13	0.15	0.15
10	2qb4	holo	33.92	15.65	0.46	0.07	0.76	0.11	0.15	0.15
10	2qei	holo	58.24	22.44	0.50	0.07	0.76	0.12	0.13	0.13
11	1hny	apo	113.28	15.34	0.36	0.24	0.70	0.12	0.19	0.19
11	1u33	holo	130.40	17.60	0.34	0.24	0.70	0.14	0.18	0.18
11	1xd0	holo	130.72	17.17	0.34	0.23	0.68	0.15	0.18	0.18
11	4gqr	holo	88.96	15.92	0.37	0.20	0.70	0.13	0.19	0.19
12	1od0	apo	117.92	20.75	0.39	0.14	0.70	0.15	0.16	0.16
12	2cet	holo	118.56	19.39	0.39	0.15	0.71	0.15	0.15	0.15
12	2j78	holo	132.96	19.47	0.39	0.16	0.69	0.15	0.17	0.17
12	2wbg	holo	130.40	19.94	0.38	0.16	0.70	0.16	0.15	0.15
13	1hl8	apo	94.08	17.25	0.39	0.18	0.75	0.14	0.14	0.14
13	2zwz	holo	53.44	14.04	0.43	0.12	0.74	0.15	0.15	0.15
13	2zx6	holo	50.88	14.93	0.42	0.12	0.76	0.13	0.15	0.15
13	2zxd	holo	84.80	28.73	0.46	0.07	0.70	0.15	0.18	0.18
14	3tpj	apo	129.44	16.94	0.38	0.17	0.71	0.19	0.16	0.16
14	3udh	holo	262.88	37.97	0.45	0.13	0.70	0.18	0.15	0.15
14	4djv	holo	136.48	17.43	0.41	0.19	0.73	0.15	0.15	0.15
14	4gid	holo	159.36	33.78	0.47	0.08	0.68	0.18	0.18	0.18
15	2o15	apo	109.44	30.56	0.45	0.06	0.64	0.17	0.22	0.22
15	2o0b	holo	76.16	25.13	0.53	0.05	0.65	0.16	0.23	0.23
15	2o0e	holo	29.76	24.10	0.55	0.02	0.65	0.15	0.23	0.23
15	2o0x	holo	6.88	17.90	0.52	0.01	0.66	0.14	0.25	0.25
16	7nn9	apo	45.44	14.74	0.46	0.10	0.65	0.22	0.16	0.16
16	1f8b	holo	101.28	22.55	0.48	0.10	0.63	0.20	0.20	0.20
16	1f8c	holo	45.76	14.83	0.44	0.11	0.65	0.22	0.16	0.16
16	1f8d	holo	107.68	22.62	0.47	0.11	0.63	0.21	0.18	0.18
17	1pud	apo	172.64	21.26	0.45	0.14	0.69	0.18	0.17	0.17
17	1n2v	holo	53.60	14.63	0.41	0.09	0.70	0.15	0.18	0.18
17	1r5y	holo	164.80	34.21	0.44	0.11	0.65	0.19	0.19	0.19
17	3ge7	holo	59.04	13.18	0.42	0.09	0.66	0.18	0.20	0.20
18	1r39	apo	199.52	27.29	0.40	0.19	0.73	0.13	0.15	0.15
18	3e93	holo	114.08	23.12	0.43	0.09	0.75	0.15	0.11	0.11
18	3gcs	holo	235.20	30.68	0.41	0.16	0.66	0.19	0.16	0.16
18	3huc	holo	88.32	17.37	0.39	0.16	0.69	0.19	0.14	0.14
19	2gfc	apo	91.36	24.36	0.48	0.08	0.71	0.17	0.14	0.14
19	1q8t	holo	66.08	20.47	0.51	0.05	0.72	0.17	0.14	0.14
19	1q8u	holo	110.56	20.82	0.49	0.08	0.72	0.17	0.13	0.13
19	3zo4	holo	78.08	20.74	0.50	0.06	0.73	0.17	0.13	0.13
20	3fwq	apo	98.40	30.37	0.46	0.08	0.68	0.15	0.18	0.18
20	2zjw	holo	103.04	26.70	0.42	0.12	0.76	0.13	0.13	0.13
20	3owj	holo	240.00	33.76	0.38	0.17	0.70	0.13	0.19	0.19
20	3pe2	holo	148.00	31.41	0.42	0.13	0.73	0.14	0.14	0.14

Table A1.4 *(continued)*. DoGSite descriptors for the extended PDBbind core set. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
21	3urj	apo	192.96	22.00	0.37	0.18	0.72	0.19	0.13	0.13
21	1gvt	holo	175.84	21.78	0.39	0.16	0.72	0.19	0.14	0.14
21	2v00	holo	182.56	22.45	0.39	0.16	0.72	0.20	0.14	0.14
21	3pww	holo	193.76	25.03	0.42	0.16	0.70	0.20	0.14	0.14
22	2zav	apo	46.56	16.15	0.48	0.09	0.61	0.20	0.22	0.22
22	3f80	holo	42.08	16.56	0.52	0.08	0.59	0.20	0.24	0.24
22	3kv2	holo	41.60	16.26	0.51	0.08	0.58	0.22	0.24	0.24
22	3mfv	holo	45.28	15.15	0.49	0.09	0.60	0.20	0.23	0.23
23	1oes	apo	2.56	14.78	0.56	0.01	0.67	0.16	0.18	0.18
23	2hb1	holo	47.20	12.83	0.42	0.14	0.68	0.10	0.25	0.25
23	2qbp	holo	47.36	12.19	0.40	0.15	0.67	0.12	0.24	0.24
23	2qbr	holo	42.72	12.76	0.42	0.13	0.68	0.13	0.22	0.22
24	3t2h	apo	104.00	21.16	0.41	0.17	0.66	0.21	0.15	0.15
24	1os0	holo	78.56	22.60	0.43	0.12	0.64	0.21	0.16	0.16
24	3fcq	holo	88.48	21.47	0.43	0.14	0.67	0.21	0.13	0.13
24	4tmn	holo	84.48	26.42	0.47	0.11	0.64	0.21	0.17	0.17
25	3pxr	apo	301.28	30.93	0.46	0.11	0.70	0.16	0.16	0.16
25	2fvd	holo	218.40	30.59	0.42	0.15	0.73	0.13	0.13	0.13
25	2r3q	holo	236.32	36.57	0.44	0.11	0.71	0.16	0.15	0.15
25	2xnb	holo	112.32	24.85	0.40	0.12	0.73	0.15	0.12	0.12
26	1qhz	apo	137.28	17.10	0.38	0.21	0.66	0.22	0.16	0.16
26	1qi0	holo	85.76	13.84	0.42	0.19	0.70	0.17	0.16	0.16
26	1w3k	holo	93.92	20.51	0.41	0.14	0.68	0.19	0.16	0.16
26	1w3l	holo	97.28	20.88	0.40	0.15	0.70	0.18	0.14	0.14
27	3cov	apo	37.76	16.60	0.43	0.05	0.68	0.15	0.19	0.19
27	3coy	holo	192.32	24.83	0.44	0.10	0.71	0.15	0.16	0.16
27	3imc	holo	122.56	24.19	0.44	0.06	0.70	0.15	0.18	0.18
27	3iub	holo	179.20	23.74	0.44	0.11	0.69	0.15	0.18	0.18
28	1rtq	apo	45.60	13.94	0.42	0.11	0.68	0.21	0.14	0.14
28	3b3s	holo	45.92	16.05	0.44	0.11	0.70	0.20	0.12	0.12
28	3b3w	holo	45.76	13.27	0.41	0.12	0.72	0.20	0.11	0.11
28	3vh9	holo	47.20	13.89	0.42	0.13	0.70	0.23	0.11	0.11
29	3kfa	apo	41.44	13.22	0.43	0.13	0.74	0.17	0.12	0.12
29	1opj	holo	131.36	21.54	0.41	0.17	0.78	0.13	0.10	0.10
29	3k5v	holo	118.72	22.69	0.43	0.14	0.78	0.13	0.10	0.10
29	3mss	holo	47.52	13.08	0.44	0.12	0.78	0.14	0.10	0.10
30	1ia8	apo	95.20	14.56	0.41	0.12	0.67	0.18	0.16	0.16
30	1nvq	holo	94.88	17.18	0.39	0.14	0.68	0.20	0.15	0.15
30	2brb	holo	101.60	18.12	0.41	0.12	0.69	0.17	0.15	0.15
30	2yex	holo	104.64	25.06	0.45	0.10	0.66	0.19	0.16	0.16
31	2zco	apo	178.88	26.84	0.40	0.12	0.72	0.16	0.15	0.15
31	2zcq	holo	137.12	28.99	0.44	0.08	0.71	0.18	0.14	0.14
31	2zcr	holo	313.44	38.08	0.40	0.15	0.71	0.17	0.15	0.15
31	3acw	holo	143.04	24.87	0.42	0.10	0.73	0.15	0.14	0.14
32	2uuf	apo	213.12	33.38	0.45	0.13	0.66	0.18	0.17	0.17
32	1bcu	holo	229.76	28.47	0.43	0.15	0.67	0.17	0.18	0.18
32	1oyt	holo	223.84	34.36	0.42	0.15	0.67	0.17	0.18	0.18
32	3utu	holo	89.12	23.76	0.46	0.09	0.67	0.18	0.17	0.17

Table A1.4 *(continued)*. DoGSite descriptors for the extended PDBbind core set. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
33	2zk0	apo	113.28	36.38	0.45	0.05	0.71	0.16	0.14	0.14
33	3b1m	holo	180.48	31.29	0.44	0.09	0.74	0.15	0.13	0.13
33	3u9q	holo	174.88	32.84	0.42	0.08	0.74	0.15	0.13	0.13
33	3v9t	holo	212.96	37.28	0.47	0.09	0.76	0.13	0.12	0.12
34	2bfx	apo	192.00	29.89	0.38	0.16	0.71	0.16	0.14	0.14
34	2vgo	holo	177.60	21.01	0.35	0.17	0.73	0.15	0.12	0.12
34	2vgp	holo	163.04	20.57	0.35	0.18	0.74	0.15	0.11	0.11
34	4b8m	holo	161.76	23.79	0.35	0.16	0.74	0.14	0.12	0.12
35	3oy5	apo	134.88	26.74	0.47	0.16	0.63	0.19	0.22	0.22
35	1o5b	holo	78.08	23.41	0.49	0.13	0.64	0.18	0.21	0.21
35	1sqa	holo	38.72	22.70	0.60	0.05	0.65	0.19	0.19	0.19
35	3kgp	holo	141.60	30.78	0.45	0.17	0.65	0.19	0.20	0.20
36	1ca2	apo	35.36	13.23	0.47	0.08	0.66	0.15	0.22	0.22
36	2weg	holo	48.80	11.87	0.42	0.11	0.67	0.16	0.20	0.20
36	3dd0	holo	44.48	13.11	0.44	0.10	0.66	0.16	0.22	0.22
36	3kwa	holo	42.40	13.47	0.47	0.10	0.69	0.12	0.23	0.23
37	3t0h	apo	128.64	24.30	0.41	0.16	0.66	0.17	0.20	0.20
37	1yc1	holo	141.28	16.80	0.40	0.15	0.66	0.18	0.19	0.19
37	2xdl	holo	133.76	23.57	0.40	0.16	0.67	0.17	0.19	0.19
37	2yki	holo	131.36	25.97	0.41	0.12	0.72	0.14	0.16	0.16
38	1mxv	apo	102.08	22.44	0.47	0.10	0.69	0.18	0.20	0.20
38	1p1q	holo	68.00	21.44	0.50	0.08	0.68	0.19	0.21	0.21
38	3bfu	holo	89.76	23.49	0.50	0.09	0.68	0.19	0.19	0.19
38	4g8m	holo	97.44	19.77	0.47	0.10	0.71	0.17	0.18	0.18
39	2p74	apo	129.28	16.61	0.37	0.23	0.60	0.26	0.23	0.23
39	3g2z	holo	138.56	16.30	0.34	0.25	0.60	0.26	0.23	0.23
39	4de1	holo	144.32	16.14	0.34	0.27	0.57	0.27	0.26	0.26
39	4de2	holo	144.48	16.03	0.33	0.25	0.57	0.28	0.26	0.26
40	4u2r	apo	89.76	16.46	0.39	0.14	0.71	0.18	0.19	0.19
40	3o6g	holo	112.48	22.66	0.49	0.10	0.70	0.17	0.18	0.18
40	4u1z	holo	255.68	29.48	0.38	0.18	0.70	0.17	0.18	0.18
40	4u21	holo	119.84	22.75	0.47	0.11	0.69	0.18	0.19	0.19
41	1hcg	apo	61.28	18.80	0.50	0.08	0.63	0.18	0.21	0.21
41	1mq6	holo	51.52	19.99	0.55	0.06	0.66	0.17	0.19	0.19
41	2xbv	holo	68.96	21.28	0.52	0.08	0.66	0.18	0.18	0.18
41	2y5h	holo	75.68	20.72	0.52	0.10	0.64	0.19	0.20	0.20
42	1dv7	apo	132.64	13.83	0.35	0.25	0.69	0.14	0.22	0.22
42	1lol	holo	132.00	16.05	0.34	0.26	0.73	0.13	0.18	0.18
42	1loq	holo	62.24	12.81	0.41	0.11	0.69	0.14	0.20	0.20
42	1lor	holo	68.16	12.61	0.40	0.12	0.70	0.14	0.19	0.19
43	2g55	apo	19.36	9.84	0.49	0.07	0.56	0.26	0.24	0.24
43	1o3f	holo	37.92	19.92	0.60	0.06	0.68	0.17	0.18	0.18
43	1uto	holo	47.20	10.80	0.43	0.17	0.56	0.27	0.24	0.24
43	3gy4	holo	33.12	10.15	0.48	0.11	0.62	0.22	0.22	0.22
44	3c0e	apo	136.32	19.48	0.41	0.14	0.68	0.19	0.16	0.16
44	2iwx	holo	262.88	29.08	0.46	0.14	0.69	0.17	0.16	0.16
44	2vw5	holo	184.96	21.80	0.41	0.15	0.67	0.18	0.18	0.18
44	2yge	holo	150.24	14.65	0.36	0.20	0.68	0.18	0.17	0.17

Table A1.4 *(continued)*. DoGSite descriptors for the extended PDBbind core set. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
45	2y7y	apo	57.28	14.25	0.49	0.14	0.71	0.18	0.14	0.14
45	2x00	holo	61.60	15.25	0.49	0.15	0.71	0.19	0.14	0.14
45	2xys	holo	63.36	16.13	0.45	0.15	0.74	0.16	0.12	0.12
45	2ymd	holo	52.64	16.27	0.60	0.10	0.71	0.19	0.15	0.15
46	1q9k	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
46	1kel	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
46	2r23	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
46	3bpc	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
47	4p7j	apo	118.88	18.86	0.35	0.17	0.67	0.19	0.17	0.17
47	3nw9	holo	66.40	19.68	0.46	0.08	0.69	0.15	0.17	0.17
47	3oe5	holo	88.64	20.94	0.44	0.10	0.69	0.15	0.18	0.18
47	3ozt	holo	76.64	19.48	0.45	0.09	0.68	0.16	0.18	0.18
48	4isv	apo	147.68	20.26	0.38	0.26	0.64	0.20	0.21	0.21
48	1igj	holo	101.92	13.99	0.38	0.16	0.69	0.18	0.19	0.19
48	1zea	holo	105.76	14.05	0.41	0.25	0.69	0.14	0.20	0.20
48	2pcp	holo	35.04	15.02	0.43	0.07	0.72	0.16	0.16	0.16
49	3hkr	apo	114.40	21.47	0.49	0.12	0.72	0.16	0.14	0.14
49	10gs	holo	182.56	24.17	0.48	0.14	0.71	0.16	0.16	0.16
49	1lbk	holo	170.24	28.24	0.45	0.13	0.71	0.16	0.15	0.15
49	2gss	holo	226.72	28.34	0.44	0.17	0.72	0.14	0.16	0.16
50	3rc6	apo	129.76	13.56	0.35	0.27	0.70	0.12	0.18	0.18
50	3su2	holo	92.00	13.34	0.37	0.21	0.72	0.13	0.15	0.15
50	3su3	holo	85.44	11.83	0.37	0.20	0.72	0.12	0.16	0.16
50	3su5	holo	85.28	12.52	0.36	0.21	0.72	0.13	0.14	0.14
51	1h05	apo	59.68	20.07	0.51	0.08	0.67	0.14	0.23	0.23
51	2xb8	holo	75.36	22.59	0.42	0.11	0.63	0.17	0.23	0.23
51	3n7a	holo	82.88	25.52	0.45	0.11	0.65	0.16	0.23	0.23
51	3n86	holo	72.48	26.22	0.45	0.11	0.65	0.16	0.24	0.24
52	3l3u	apo	51.84	12.56	0.39	0.13	0.75	0.17	0.11	0.11
52	3zso	holo	99.68	11.26	0.31	0.25	0.74	0.15	0.12	0.12
52	3zsq	holo	99.04	12.27	0.32	0.24	0.74	0.15	0.13	0.13
52	3zsx	holo	99.20	11.77	0.31	0.25	0.75	0.17	0.10	0.10
53	4lzu	apo	1.28	18.82	0.69	0.00	0.82	0.10	0.08	0.08
53	3nq3	holo	123.36	23.19	0.42	0.12	0.78	0.13	0.10	0.10
53	3ueu	holo	128.16	22.57	0.44	0.12	0.81	0.11	0.09	0.09
53	3uex	holo	104.16	22.58	0.42	0.12	0.79	0.13	0.10	0.10
54	2oxu	apo	105.44	23.33	0.40	0.14	0.71	0.16	0.15	0.15
54	3ehy	holo	98.72	21.08	0.39	0.14	0.68	0.18	0.17	0.17
54	3f17	holo	114.08	22.42	0.36	0.16	0.72	0.14	0.15	0.15
54	3lka	holo	117.76	22.16	0.37	0.17	0.69	0.16	0.16	0.16
55	3cfm	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
55	3cft	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
55	4des	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
55	4dew	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
56	1kf3	apo	153.92	23.06	0.36	0.24	0.65	0.17	0.20	0.20
56	1u1b	holo	138.40	20.59	0.32	0.29	0.63	0.16	0.23	0.23
56	1w4o	holo	130.88	19.41	0.33	0.29	0.64	0.17	0.22	0.22
56	3dxg	holo	167.04	21.03	0.36	0.26	0.65	0.16	0.21	0.21

Table A1.4 *(continued)*. DoGSite descriptors for the extended PDBbind core set. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
57	2h46	apo	184.96	21.25	0.43	0.21	0.71	0.13	0.19	0.19
57	1jyq	holo	58.40	17.07	0.51	0.13	0.65	0.17	0.21	0.21
57	3ov1	holo	66.88	16.82	0.51	0.14	0.67	0.15	0.21	0.21
57	3s8o	holo	68.64	13.18	0.46	0.18	0.61	0.19	0.24	0.24
58	2g69	apo	85.76	13.08	0.41	0.25	0.76	0.10	0.16	0.16
58	1a30	holo	100.80	11.24	0.33	0.26	0.68	0.17	0.17	0.17
58	3cyx	holo	97.12	10.73	0.33	0.25	0.68	0.19	0.15	0.15
58	4djr	holo	105.44	13.78	0.33	0.27	0.72	0.16	0.14	0.14
59	1dp0	apo	16.96	15.96	0.64	0.09	0.66	0.18	0.18	0.18
59	3i3b	holo	19.52	18.12	0.62	0.09	0.64	0.20	0.18	0.18
59	3muz	holo	14.88	21.69	0.73	0.05	0.66	0.17	0.19	0.19
59	3vd4	holo	14.24	16.20	0.66	0.07	0.65	0.21	0.16	0.16
60	4jho	apo	108.96	22.10	0.40	0.10	0.69	0.16	0.19	0.19
60	4jie	holo	102.56	17.58	0.37	0.11	0.71	0.13	0.19	0.19
60	4re2	holo	92.80	21.90	0.39	0.09	0.68	0.17	0.19	0.19
60	4re4	holo	90.40	21.27	0.40	0.09	0.69	0.16	0.18	0.18
61	2bgr	apo	20.80	10.50	0.53	0.13	0.73	0.18	0.18	0.18
61	1n1m	holo	46.72	15.00	0.45	0.17	0.71	0.18	0.18	0.18
61	2ole	holo	74.56	25.07	0.55	0.11	0.69	0.16	0.17	0.17
61	3nox	holo	23.84	13.03	0.54	0.12	0.74	0.18	0.16	0.16
62	3kpj	apo	50.72	34.12	0.52	0.03	0.72	0.15	0.15	0.15
62	1hnn	holo	22.40	34.85	0.55	0.01	0.70	0.17	0.16	0.16
62	2g70	holo	26.08	35.28	0.51	0.02	0.73	0.15	0.15	0.15
62	2obf	holo	43.68	29.89	0.50	0.03	0.72	0.17	0.14	0.14
63	2am9	apo	36.16	9.31	0.43	0.23	0.74	0.18	0.10	0.10
63	2pip	holo	38.72	8.14	0.43	0.24	0.78	0.17	0.08	0.08
63	2pit	holo	57.28	12.43	0.46	0.20	0.74	0.19	0.09	0.09
63	2piv	holo	37.12	7.81	0.41	0.26	0.73	0.21	0.09	0.09
64	1qia	apo	78.72	20.43	0.38	0.13	0.69	0.16	0.15	0.15
64	1hfs	holo	122.88	25.06	0.40	0.14	0.66	0.17	0.18	0.18
64	1sln	holo	163.68	27.98	0.38	0.18	0.70	0.14	0.17	0.17
64	2d1o	holo	199.84	28.72	0.32	0.23	0.70	0.15	0.16	0.16
65	1ous	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
65	2jdm	holo	153.12	19.93	0.41	0.29	0.57	0.29	0.18	0.18
65	2jdu	holo	61.28	12.17	0.43	0.28	0.52	0.34	0.20	0.20
65	2jdy	holo	64.48	11.74	0.41	0.29	0.52	0.32	0.20	0.20

Table A1.5. The dataset of cryptic site pairs. The characteristics of all structures of the set of cryptic site pairs are shown together with the affinity/activity of the ligands. Activity or affinity data for the ligands was extracted from the PDBbind database[215]. If this information was extracted from other databases, the respective database (Binding MOAD[218] or BindingDB[219]) is given in brackets.

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity
1	1a8i	A	apo	P00489	2.4.1.1				
1	2ieg	B	holo	P00489	2.4.1.1	FRY	902	B	IC_{50}=0.135 µM (Binding MOAD)
2	1ade	A	apo	P0A7D4	6.3.4.4				
2	1cib	A	holo	P0A7D4	6.3.4.4	IMP	440	A	n/a

Table A1.5 *(continued)*. The dataset of cryptic site pairs. The characteristics of all structures of the set of cryptic site pairs are shown together with the affinity/activity of the ligands. Activity or affinity data for the ligands was extracted from the PDBbind database[215]. If this information was extracted from other databases, the respective database (Binding MOAD[218] or BindingDB[219]) is given in brackets.

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity
3	1alb	A	apo	P04117					
3	1lic	A	holo	P04117		HDS	133	A	n/a
4	1alv	A	apo	P04574					
4	1nx3	A	holo	P04574		ISA	267	A	n/a
5	1b6b	A	apo	Q29495	2.3.1.87				
5	1kuv	A	holo	Q29495	2.3.1.87	CA5	901	A	K_i=22 nM (Binding MOAD)
6	1bnc	B	apo	P24182	6.3.4.14				
6	2v5a	A	holo	P24182	6.3.4.14	LZL	1447	A	IC_{50}=150 nM
7	1bsq	A	apo	P02754					
7	1gx8	A	holo	P02754		RTL	1163	A	K_d=0.43 µM
8	1dub	D	apo	P14604	4.2.1.17				
8	1ey3	F	holo	P14604	4.2.1.17	DAK	505	F	n/a
9	1e2x	A	apo	P0A8V6					
9	1h9g	A	holo	P0A8V6		MYR	1229	A	K_d=59 nM (Binding MOAD)
10	1ecj	D	apo	P0AG16	2.4.2.14				
10	1ecc	B	holo	P0AG16	2.4.2.14	PCP	505	B	n/a
11	1ex6	A	apo	P15454	2.7.4.8				
11	1gky	A	holo	P15454	2.7.4.8	5GP	187	A	n/a
12	1exm	A	apo	Q5SHN6	3.1.5.1				
12	1ha3	B	holo	Q5SHN6	3.1.5.1	MAU	408	B	n/a
13	1fa9	A	apo	P06737	2.4.1.1				
13	1l5s	B	holo	P06737	2.4.1.1	URC	1863	B	K_d=550 µM
14	1fxx	A	apo	P04995	3.1.11.1				
14	3hl8	A	holo	P04995	3.1.11.1	BBP	501	A	K_i=32 µM
15	1g24	D	apo	P15879	2.4.2.-				
15	1gzf	C	holo	P15879	2.4.2.-	NIR	1252	C	n/a
16	1g4e	B	apo	P39594	2.5.1.3				
16	1g67	B	holo	P39594	2.5.1.3	POP, TZP	2004, 2006	B, B	n/a
17	1h09	A	apo	P15057	3.2.1.17				
17	2ixu	A	holo	P15057	3.2.1.17	MU2	1341	A	n/a
18	1hag	E	apo	P00734	3.4.21.5				
18	1ghy	H	holo	P00734	3.4.21.5	121	246	H	K_i=0.008 µM
19	1hka	A	apo	P26281	2.7.6.3				
19	3ip0	A	holo	P26281	2.7.6.3	HHS	182	A	n/a
20	1imf	A	apo	P29218	3.1.3.25				
20	1imb	B	holo	P29218	3.1.3.25	LIP	279	B	n/a
21	1jbu	H	apo	P08709	3.4.21.21				
21	1wun	H	holo	P08709	3.4.21.21	P5B	2001	H	IC_{50}=62 nM
22	1jwp	A	apo	P62593	3.5.2.6				
22	1pzo	A	holo	P62593	3.5.2.6	CBT, CBT	300, 301	A, A	K_i=480 µM
23	1k3f	B	apo	P12758	2.4.2.3				
23	1u1d	F	holo	P12758	2.4.2.3	181	8400	F	n/a
24	1k5h	C	apo	P45568	1.1.1.267				
24	2egh	B	holo	P45568	1.1.1.267	FOM	2001	B	K_i=38 nM (Binding MOAD)

Table A1.5 *(continued)*. The dataset of cryptic site pairs. The characteristics of all structures of the set of cryptic site pairs are shown together with the affinity/activity of the ligands. Activity or affinity data for the ligands was extracted from the PDBbind database[215]. If this information was extracted from other databases, the respective database (Binding MOAD[218] or BindingDB[219]) is given in brackets.

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity
25	1ks9	A	apo	P0A9J4	1.1.1.169				
25	2ofp	A	holo	P0A9J4	1.1.1.169	PAF	602	A	Kd=160 µM (Binding MOAD)
26	1kz7	D	apo	P60953					
26	1grn	A	holo	P60953		AF3	200	A	n/a
27	1my0	B	apo	P19491					
27	1n0t	D	holo	P19491		AT1	266	D	IC_{50}=12.2 µM
28	1my1	C	apo	P19491					
28	1ftl	A	holo	P19491		DNQ	365	A	IC_{50}=998 nM
29	1nep	A	apo	P79345					
29	2hka	C	holo	P79345		C3S	401	C	n/a
30	1ni6	D	apo	P09601	1.14.99.3				
30	3hok	B	holo	P09601	1.14.99.3	Q80	301	B	IC_{50}=2.1 µM (Binding MOAD)
31	1nuw	A	apo	P00636	3.1.3.11				
31	1eyj	B	holo	P00636	3.1.3.11	AMP	342	B	IC_{50}=1.3 µM (BindingDB)
32	1pkl	B	apo	Q27686	2.7.1.40				
32	3hqp	P	holo	Q27686	2.7.1.40	ATP, FDP, OXL	1001, 510, 700	P, P, P	n/a
33	1pzt	A	apo	P08037	2.4.1.90				
33	1pzy	D	holo	P08037	2.4.1.22, 2.4.1.90, 2.4.1.38	UDP	409	D	n/a
34	1qlw	B	apo	Q7SIA5					
34	2wkw	B	holo	Q7SIA5		W22	577	B	n/a
35	1r1w	A	apo	P08581	2.7.10.1				
35	3f82	A	holo	P08581	2.7.10.1	353	2001	A	K_i=4.6 nM
36	1rdw	X	apo	P68135					
36	1j6z	A	holo	P68135		RHO	1381	A	n/a
37	1rhb	A	apo	P61823	3.1.27.5				
37	2w5k	B	holo	P61823	3.1.27.5	NDP	1125	B	K_i=12 µM (Binding MOAD)
38	1rtc	A	apo	P02879	3.2.2.22				
38	1br6	A	holo	P02879	3.2.2.22	PT1	301	A	K_i=0.6 mM
39	1swx	A	apo	Q9NZD2					
39	2eum	A	holo	Q9NZD2		LAT	300	A	K_d=0.2 µM (Binding MOAD)
40	1tqo	A	apo	P00644	3.1.31.1				
40	1tr5	A	holo	P00644	3.1.31.1	THP	300	A	n/a
41	1uk2	A	apo	P0C6X7	3.4.24.-				
41	2gz7	A	holo	Q6RCZ9	3.4.22.-	D3F	307	A	K_d=0.2 µM
42	1w50	A	apo	P56817	3.4.23.46				
42	3ixj	C	holo	P56817	3.4.23.46	586	500	C	IC_{50}=0.32 nM
43	1xcg	B	apo	P61586					
43	1ow3	B	holo	P61586		GDP	680	B	n/a
44	1z92	A	apo	P60568					
44	1py2	A	holo	P60568		FRH	201	A	IC_{50}=0.06 µM

Table A1.5 *(continued)*. The dataset of cryptic site pairs. The characteristics of all structures of the set of cryptic site pairs are shown together with the affinity/activity of the ligands. Activity or affinity data for the ligands was extracted from the PDBbind database[215]. If this information was extracted from other databases, the respective database (Binding MOAD[218] or BindingDB[219]) is given in brackets.

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity
45	1zah	B	apo	P00883	4.1.2.13				
45	2ot1	D	holo	P00883	4.1.2.13	N3P	3004	D	K_i=0.1 mM
46	2air	H	apo	P0A7F3	2.1.3.2				
46	1za1	D	holo	P0A7F3	2.1.3.2	CTP	902	D	n/a
47	2aka	A	apo	P08799					
47	1yv3	A	holo	P08799		BIT	800	A	n/a
48	2ax9	A	apo	P10275					
48	2piq	A	holo	P10275		RB1	933	A	n/a
49	2bf3	A	apo	Q00459	1.14.13.-				
49	3dhh	E	holo	Q00459	1.14.13.-	BML	507	E	n/a
50	2bls	B	apo	P00811	3.5.2.6				
50	3gqz	A	holo	P00811	3.5.2.6	GF7	362	A	K_i=6.7mM
51	2brk	A	apo	P26663	2.7.7.48				
51	2gir	B	holo	P26663	2.7.7.48	NN3	2001	B	IC_{50}=140 nM (Binding MOAD)
52	2bu8	A	apo	Q15119	2.7.1.99				
52	2bu2	A	holo	Q15119	2.7.1.99	TF1	1387	A	n/a
53	2cga	B	apo	P00766	3.4.21.1				
53	1afq	C	holo	P00766	3.4.21.1	0FG	301	C	n/a
54	2cm2	A	apo	P18031	3.1.3.48				
54	2h4k	A	holo	P18031	3.1.3.48	509	300	A	K_i=3.2 µM
55	2gfc	A	apo	P00517	2.7.11.11				
55	2jds	A	holo	P00517	2.7.11.11	L20	1351	A	K_i=6.3 nM
56	2h4e	B	apo	P02766					
56	3cfn	B	holo	P02766		2AN	1	B	K_d=10 µM
57	2iyt	A	apo	P9WPY3	2.7.1.71				
57	2iyq	A	holo	P9WPY3	2.7.1.71	ADP, SKM	201,202	A, A	n/a
58	2ohg	A	apo	Q9A1B7	5.1.1.3				
58	2ohv	A	holo	Q9A1B7	5.1.1.3	NHL	400	A	IC_{50}=0.35 µM
59	2q8f	A	apo	Q15118	2.7.11.2				
59	2q8h	A	holo	Q15118	2.7.11.2	TF4	438	A	K_d=1 mM
60	2qfo	B	apo	P07900					
60	2wi7	A	holo	P07900		2KL	1224	A	IC_{50}=58 nM
61	2qlr	C	apo	Q8N5Z0	2.6.1.7, 2.6.1.39				
61	3dc1	A	holo	Q8N5Z0	2.6.1.7, 2.6.1.39	AKG	426	A	n/a
62	2wgb	A	apo	Q58L87					
62	2v57	A	holo	Q58L87		PRL	1188	A	K_d=79 nM
63	2wgq	B	apo	P46883	1.4.3.6				
63	1d6y	B	holo	P46883	1.4.3.6	HY1	3001	B	n/a
64	2yqc	A	apo	O74933	2.7.7.23				
64	2yqs	A	holo	O74933	2.7.7.23	UD1	1001	A	n/a
65	2zb1	A	apo	Q16539	2.7.11.24				
65	2npq	A	holo	Q16539	2.7.11.24	BOG, BOG	1000, 2000	A, A	K_d=3.08 µM (Binding MOAD)
66	3b7d	E	apo	P19491					
66	2al4	F	holo	P19491		CX6	803	F	n/a

Table A1.5 *(continued)*. The dataset of cryptic site pairs. The characteristics of all structures of the set of cryptic site pairs are shown together with the affinity/activity of the ligands. Activity or affinity data for the ligands was extracted from the PDBbind database[215]. If this information was extracted from other databases, the respective database (Binding MOAD[218] or BindingDB[219]) is given in brackets.

Group	PDB-ID	Chain-ID	State	UniProt-Accession	EC number	Ligand-Name	Ligand-ID	Ligand-Chain	Ligand affinity/ activity
67	3bl9	B	apo	Q96C86	3.-.-.-				
67	3bl7	A	holo	Q96C86	3.-.-.-	DD1	901	A	IC_{50}=4 nM
68	3che	A	apo	Q873X9	3.2.1.14				
68	2iuz	B	holo	Q873X9	3.2.1.14	D1H	1438	B	K_i=2.8 µM
69	3dxn	A	apo	Q3HNM6	2.7.11.17				
69	3hzt	A	holo	Q3HNM6	2.7.11.17	J60	540	A	n/a
70	3fdl	A	apo	Q07817					
70	2yxj	A	holo	Q07817		N3C	1001	A	K_i=0.5 nM
71	3gxd	B	apo	P04062	3.2.1.45				
71	2wcg	A	holo	P04062	3.2.1.45	MT5	1509	A	K_i=0.42 µM
72	3h5r	A	apo	Q47506					
72	3h9j	D	holo	Q47506		APC	359	D	n/a
73	3hqd	A	apo	P52732					
73	1q0b	B	holo	P52732		NAT	605	B	IC_{50}=6.7 µM
74	3kqa	B	apo	P33038	2.5.1.7				
74	3lth	A	holo	P33038	2.5.1.7	UD1	450	A	n/a
75	3l7u	C	apo	P15531	2.7.4.6				
75	2hvd	C	holo	P15531	2.7.4.6	ADP	160	C	n/a
76	3mn9	A	apo	P10987					
76	3eks	A	holo	P10987		CY9	903	A	K_d=2 µM
77	3nnu	A	apo	Q16539	2.7.11.24				
77	3hl7	A	holo	Q16539	2.7.11.24	I46	402	A	n/a
78	3peo	G	apo	Q8WSF8					
78	2bys	J	holo	Q8WSF8		LOB	301	J	K_d=0.3 nM
79	3puw	E	apo	P0AEX9					
79	1fqc	A	holo	P0AEX9		GLO	1000	A	n/a
80	4ake	B	apo	P69441	2.7.4.3				
80	1ank	B	holo	P69441	2.7.4.3	ANP	216	B	n/a
81	4hb2	C	apo	P30822					
81	4hat	C	holo	P30822		LMB	1101	C	n/a

Table A1.6. RMSD values between ligand-bound and unbound structures for the cryptic site pairs. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
1	1a8i	A	apo	0.46	0.36	0.39
1	2ieg	B	holo	0.46	0.36	0.39
2	1ade	A	apo	1.99	1.67	1.16
2	1cib	A	holo	1.99	1.67	1.16
3	1alb	A	apo	0.52	2.35	0.45
3	1lic	A	holo	0.52	2.35	0.45
4	1alv	A	apo	0.63	0.72	0.35
4	1nx3	A	holo	0.63	0.72	0.35
5	1b6b	A	apo	2.25	1.45	0.48
5	1kuv	A	holo	2.25	1.45	0.48

Table A1.6 *(continued)*. RMSD values between ligand-bound and unbound structures for the cryptic site pairs. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
6	1bnc	B	apo	2.16	2.62	0.76
6	2v5a	A	holo	2.16	2.62	0.76
7	1bsq	A	apo	1.11	1.06	0.27
7	1gx8	A	holo	1.11	1.06	0.27
8	1dub	D	apo	0.60	1.24	0.52
8	1ey3	F	holo	0.60	1.24	0.52
9	1e2x	A	apo	1.66	2.21	0.96
9	1h9g	A	holo	1.66	2.21	0.96
10	1ecj	D	apo	1.47	1.24	0.96
10	1ecc	B	holo	1.47	1.24	0.96
11	1ex6	A	apo	2.99	2.11	0.62
11	1gky	A	holo	2.99	2.11	0.62
12	1exm	A	apo	2.15	2.84	0.39
12	1ha3	B	holo	2.15	2.84	0.39
13	1fa9	A	apo	1.63	1.56	0.35
13	1l5s	B	holo	1.63	1.56	0.35
14	1fxx	A	apo	1.25	1.48	0.35
14	3hl8	A	holo	1.25	1.48	0.35
15	1g24	D	apo	0.72	2.16	0.43
15	1gzf	C	holo	0.72	2.16	0.43
16	1g4e	B	apo	0.33	2.64	0.46
16	1g67	B	holo	0.33	2.64	0.46
17	1h09	A	apo	0.55	0.22	0.34
17	2ixu	A	holo	0.55	0.22	0.34
18	1hag	E	apo	1.50	2.96	0.91
18	1ghy	H	holo	1.50	2.96	0.91
19	1hka	A	apo	1.30	1.10	0.44
19	3ip0	A	holo	1.30	1.10	0.44
20	1imf	A	apo	1.06	1.53	0.39
20	1imb	B	holo	1.06	1.53	0.39
21	1jbu	H	apo	2.83	2.45	0.55
21	1wun	H	holo	2.83	2.45	0.55
22	1jwp	A	apo	0.92	3.25	0.25
22	1pzo	A	holo	0.92	3.25	0.25
23	1k3f	B	apo	1.91	1.86	0.61
23	1u1d	F	holo	1.91	1.86	0.61
24	1k5h	C	apo	1.49	0.35	0.80
24	2egh	B	holo	1.49	0.35	0.80
25	1ks9	A	apo	1.31	1.14	0.38
25	2ofp	A	holo	1.31	1.14	0.38
26	1kz7	D	apo	1.63	1.09	0.38
26	1grn	A	holo	1.63	1.09	0.38
27	1my0	B	apo	1.94	0.03	1.01
27	1n0t	D	holo	1.94	0.03	1.01
28	1my1	C	apo	1.92	2.61	0.83
28	1ftl	A	holo	1.92	2.61	0.83
29	1nep	A	apo	1.12	3.08	0.47
29	2hka	C	holo	1.12	3.08	0.47

Table A1.6 *(continued)*. RMSD values between ligand-bound and unbound structures for the cryptic site pairs. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
30	1ni6	D	apo	1.53	1.84	0.87
30	3hok	B	holo	1.53	1.84	0.87
31	1nuw	A	apo	1.30	1.17	0.65
31	1eyj	B	holo	1.30	1.17	0.65
32	1pkl	B	apo	3.63	1.43	0.79
32	3hqp	P	holo	3.63	1.43	0.79
33	1pzt	A	apo	1.58	1.27	0.30
33	1pzy	D	holo	1.58	1.27	0.30
34	1qlw	B	apo	0.43	1.64	0.28
34	2wkw	B	holo	0.43	1.64	0.28
35	1r1w	A	apo	1.74	1.16	0.60
35	3f82	A	holo	1.74	1.16	0.60
36	1rdw	X	apo	1.48	1.61	0.46
36	1j6z	A	holo	1.48	1.61	0.46
37	1rhb	A	apo	0.57	0.86	1.10
37	2w5k	B	holo	0.57	0.86	1.10
38	1rtc	A	apo	0.77	0.32	0.34
38	1br6	A	holo	0.77	0.32	0.34
39	1swx	A	apo	1.16	0.20	0.50
39	2eum	A	holo	1.16	0.20	0.50
40	1tqo	A	apo	0.64	0.09	0.55
40	1tr5	A	holo	0.64	0.09	0.55
41	1uk2	A	apo	1.19	0.64	0.43
41	2gz7	A	holo	1.19	0.64	0.43
42	1w50	A	apo	1.00	3.60	0.16
42	3ixj	C	holo	1.00	3.60	0.16
43	1xcg	B	apo	1.40	1.97	0.53
43	1ow3	B	holo	1.40	1.97	0.53
44	1z92	A	apo	1.30	1.54	0.51
44	1py2	A	holo	1.30	1.54	0.51
45	1zah	B	apo	0.24	1.44	0.30
45	2ot1	D	holo	0.24	1.44	0.30
46	2air	H	apo	1.46	1.07	0.72
46	1za1	D	holo	1.46	1.07	0.72
47	2aka	A	apo	3.52	1.56	0.71
47	1yv3	A	holo	3.52	1.56	0.71
48	2ax9	A	apo	0.38	0.09	0.48
48	2piq	A	holo	0.38	0.09	0.48
49	2bf3	A	apo	0.69	1.98	0.68
49	3dhh	E	holo	0.69	1.98	0.68
50	2bls	B	apo	0.67	1.22	0.75
50	3gqz	A	holo	0.67	1.22	0.75
51	2brk	A	apo	1.24	1.56	0.45
51	2gir	B	holo	1.24	1.56	0.45
52	2bu8	A	apo	0.70	3.54	0.38
52	2bu2	A	holo	0.70	3.54	0.38
53	2cga	B	apo	1.41	1.57	1.33
53	1afq	C	holo	1.41	1.57	1.33

Table A1.6 *(continued)*. RMSD values between ligand-bound and unbound structures for the cryptic site pairs. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
54	2cm2	A	apo	1.07	2.19	0.55
54	2h4k	A	holo	1.07	2.19	0.55
55	2gfc	A	apo	0.60	1.33	0.44
55	2jds	A	holo	0.60	1.33	0.44
56	2h4e	B	apo	1.10	1.67	0.32
56	3cfn	B	holo	1.10	1.67	0.32
57	2iyt	A	apo	1.78	1.13	0.83
57	2iyq	A	holo	1.78	1.13	0.83
58	2ohg	A	apo	0.86	2.03	0.90
58	2ohv	A	holo	0.86	2.03	0.90
59	2q8f	A	apo	0.54	0.32	0.86
59	2q8h	A	holo	0.54	0.32	0.86
60	2qfo	B	apo	1.01	2.83	0.26
60	2wi7	A	holo	1.01	2.83	0.26
61	2qlr	C	apo	0.47	0.70	0.27
61	3dc1	A	holo	0.47	0.70	0.27
62	2wgb	A	apo	3.43	1.96	1.24
62	2v57	A	holo	3.43	1.96	1.24
63	2wgq	B	apo	0.33	1.48	0.22
63	1d6y	B	holo	0.33	1.48	0.22
64	2yqc	A	apo	1.33	2.31	0.28
64	2yqs	A	holo	1.33	2.31	0.28
65	2zb1	A	apo	1.25	0.74	0.24
65	2npq	A	holo	1.25	0.74	0.24
66	3b7d	E	apo	1.65	0.52	0.22
66	2al4	F	holo	1.65	0.52	0.22
67	3bl9	B	apo	3.05	0.40	0.23
67	3bl7	A	holo	3.05	0.40	0.23
68	3che	A	apo	0.39	0.9	0.63
68	2iuz	B	holo	0.39	0.9	0.63
69	3dxn	A	apo	2.28	0.43	0.72
69	3hzt	A	holo	2.28	0.43	0.72
70	3fdl	A	apo	0.83	1.31	0.33
70	2yxj	A	holo	0.83	1.31	0.33
71	3gxd	B	apo	0.88	0.6	0.52
71	2wcg	A	holo	0.88	0.6	0.52
72	3h5r	A	apo	0.72	0.56	0.36
72	3h9j	D	holo	0.72	0.56	0.36
73	3hqd	A	apo	2.10	0.82	1.59
73	1q0b	B	holo	2.10	0.82	1.59
74	3kqa	B	apo	1.74	2.74	0.45
74	3lth	A	holo	1.74	2.74	0.45
75	3l7u	C	apo	0.74	1.86	0.34
75	2hvd	C	holo	0.74	1.86	0.34
76	3mn9	A	apo	0.87	1.75	0.51
76	3eks	A	holo	0.87	1.75	0.51
77	3nnu	A	apo	1.57	1.33	0.91
77	3hl7	A	holo	1.57	1.33	0.91

Table A1.6 *(continued)*. RMSD values between ligand-bound and unbound structures for the cryptic site pairs. TM-align[216] was applied to compare complete protein structures and protein binding sites. ProBiS[104] was applied to compare binding sites based on spatially and physicochemically similar binding site patches.

Group	PDB-ID	Chain-ID	State	TM-align $RMSD_{apo\text{-}holo}$ (complete proteins) [Å]	TM-align $RMSD_{apo\text{-}holo}$ (binding site) [Å]	ProBiS $RMSD_{apo\text{-}holo}$ [Å]
78	3peo	G	apo	1.11	3.21	0.71
78	2bys	J	holo	1.11	3.21	0.71
79	3puw	E	apo	2.87	1.94	0.41
79	1fqc	A	holo	2.87	1.94	0.41
80	4ake	B	apo	3.74	2.37	0.81
80	1ank	B	holo	3.74	2.37	0.81
81	4hb2	C	apo	0.69	2.64	0.46
81	4hat	C	holo	0.69	2.64	0.46

Table A1.7. DoGSite descriptors for the set of cryptic site pairs. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
1	1a8i	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
1	2ieg	holo	20.32	7.71	0.51	0.16	0.71	0.12	0.17	0.14
2	1ade	apo	226.88	38.97	0.38	0.15	0.64	0.18	0.21	0.02
2	1cib	holo	192.00	37.01	0.40	0.11	0.65	0.18	0.20	0.03
3	1alb	apo	68.80	20.38	0.44	0.05	0.71	0.15	0.16	0.09
3	1lic	holo	58.08	15.30	0.35	0.06	0.68	0.18	0.19	0.08
4	1alv	apo	39.36	11.79	0.51	0.14	0.80	0.10	0.10	0.27
4	1nx3	holo	32.64	11.49	0.44	0.09	0.79	0.09	0.12	0.31
5	1b6b	apo	111.04	21.57	0.41	0.08	0.74	0.14	0.13	0.13
5	1kuv	holo	195.52	27.73	0.38	0.15	0.76	0.12	0.14	0.17
6	1bnc	apo	96.32	13.46	0.36	0.26	0.68	0.18	0.14	0.19
6	2v5a	holo	72.16	19.07	0.41	0.08	0.71	0.16	0.14	0.11
7	1bsq	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
7	1gx8	holo	40.00	21.41	0.49	0.05	0.82	0.09	0.08	0.09
8	1dub	apo	244.48	22.53	0.41	0.21	0.68	0.18	0.15	0.10
8	1ey3	holo	225.28	24.14	0.39	0.19	0.68	0.18	0.15	0.07
9	1e2x	apo	55.04	21.02	0.55	0.07	0.71	0.17	0.16	0.24
9	1h9g	holo	227.52	27.36	0.46	0.11	0.72	0.15	0.16	0.18
10	1ecj	apo	122.40	17.23	0.45	0.20	0.72	0.15	0.16	0.00
10	1ecc	holo	71.84	26.38	0.47	0.06	0.68	0.17	0.19	0.09
11	1ex6	apo	261.28	20.91	0.36	0.30	0.63	0.23	0.19	0.07
11	1gky	holo	276.16	28.20	0.37	0.20	0.68	0.18	0.16	0.13
12	1exm	apo	131.68	21.43	0.39	0.22	0.65	0.21	0.18	0.10
12	1ha3	holo	451.52	41.80	0.37	0.19	0.68	0.16	0.17	0.07
13	1fa9	apo	133.76	32.23	0.50	0.12	0.65	0.17	0.20	0.08
13	1l5s	holo	17.60	12.56	0.62	0.08	0.76	0.10	0.15	0.27
14	1fxx	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
14	3hl8	holo	25.60	11.72	0.50	0.12	0.77	0.11	0.12	0.16
15	1g24	apo	203.36	24.97	0.39	0.19	0.68	0.17	0.19	0.15
15	1gzf	holo	126.72	20.43	0.41	0.13	0.65	0.18	0.21	0.13
16	1g4e	apo	64.80	11.93	0.42	0.16	0.61	0.17	0.26	0.19
16	1g67	holo	42.72	20.44	0.40	0.05	0.65	0.16	0.22	0.11

Table A1.7 *(continued)*. DoGSite descriptors for the set of cryptic site pairs. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
17	1h09	apo	44.48	7.16	0.40	0.31	0.74	0.12	0.15	0.03
17	2ixu	holo	167.84	25.18	0.43	0.22	0.71	0.17	0.14	0.25
18	1hag	apo	142.56	12.71	0.31	0.33	0.77	0.15	0.11	0.34
18	1ghy	holo	221.92	30.19	0.43	0.15	0.67	0.17	0.18	0.14
19	1hka	apo	149.28	22.97	0.40	0.19	0.67	0.18	0.16	0.09
19	3ip0	holo	266.24	26.23	0.40	0.17	0.68	0.16	0.18	0.11
20	1imf	apo	127.20	29.65	0.46	0.17	0.69	0.18	0.14	0.16
20	1imb	holo	228.00	23.87	0.40	0.16	0.65	0.21	0.18	0.08
21	1jbu	apo	231.84	20.21	0.33	0.37	0.66	0.20	0.18	0.13
21	1wun	holo	103.84	22.94	0.47	0.11	0.64	0.19	0.20	0.12
22	1jwpA	apo	89.92	13.39	0.39	0.22	0.58	0.25	0.23	0.05
22	1pzoA	holo	192.80	30.10	0.42	0.14	0.70	0.18	0.14	0.03
23	1k3f	apo	90.72	13.71	0.38	0.26	0.65	0.22	0.18	0.18
23	1u1d	holo	75.36	18.75	0.47	0.10	0.65	0.18	0.21	0.05
24	1k5h	apo	35.68	15.82	0.52	0.10	0.64	0.24	0.18	0.07
24	2egh	holo	108.48	22.29	0.45	0.10	0.66	0.19	0.18	0.09
25	1ks9	apo	162.24	32.30	0.50	0.11	0.66	0.16	0.20	0.08
25	2ofp	holo	458.72	29.34	0.35	0.24	0.67	0.16	0.19	0.04
26	1kz7	apo	113.60	15.50	0.40	0.22	0.65	0.16	0.21	0.02
26	1grn	holo	86.56	18.61	0.44	0.10	0.69	0.16	0.18	0.07
27	1my0	apo	79.84	20.40	0.51	0.08	0.69	0.18	0.19	0.10
27	1n0t	holo	259.36	35.10	0.47	0.14	0.73	0.16	0.16	0.14
28	1my1	apo	106.88	20.56	0.47	0.12	0.71	0.17	0.18	0.11
28	1ftl	holo	248.16	25.72	0.34	0.24	0.69	0.18	0.21	0.14
29	1nep	apo	43.20	12.09	0.53	0.18	0.80	0.16	0.07	0.18
29	2hka	holo	83.20	22.13	0.42	0.08	0.81	0.12	0.08	0.13
30	1ni6	apo	173.60	28.29	0.47	0.10	0.73	0.14	0.14	0.20
30	3hok	holo	383.84	32.98	0.40	0.19	0.70	0.16	0.16	0.13
31	1nuw	apo	177.12	19.47	0.41	0.20	0.64	0.18	0.20	0.05
31	1eyj	holo	154.24	28.55	0.52	0.10	0.70	0.16	0.15	0.08
32	1pkl	apo	110.88	18.57	0.44	0.13	0.66	0.20	0.16	0.08
32	3hqp	holo	116.48	28.89	0.43	0.07	0.66	0.17	0.20	0.09
33	1pzt	apo	172.96	19.07	0.43	0.16	0.67	0.16	0.18	0.17
33	1pzy	holo	140.16	21.34	0.37	0.16	0.65	0.17	0.18	0.09
34	1qlw	apo	95.04	16.09	0.42	0.15	0.70	0.14	0.19	0.21
34	2wkw	holo	143.52	17.14	0.40	0.16	0.67	0.16	0.19	0.14
35	1r1w	apo	52.80	14.07	0.43	0.10	0.71	0.15	0.15	0.09
35	3f82	holo	95.20	25.68	0.45	0.08	0.76	0.12	0.12	0.14
36	1rdw	apo	42.40	7.02	0.40	0.29	0.86	0.11	0.11	0.39
36	1j6z	holo	192.64	23.58	0.36	0.22	0.75	0.14	0.13	0.18
37	1rhb	apo	132.16	22.29	0.39	0.23	0.66	0.18	0.19	0.15
37	2w5k	holo	153.44	22.94	0.41	0.25	0.66	0.17	0.20	0.13
38	1rtc	apo	70.24	7.89	0.34	0.32	0.62	0.19	0.24	0.24
38	1br6	holo	80.96	15.41	0.43	0.14	0.69	0.18	0.15	0.14
39	1swx	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
39	2eum	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
40	1tqo	apo	55.20	12.24	0.42	0.15	0.67	0.17	0.19	0.12
40	1tr5	holo	93.60	22.10	0.45	0.13	0.66	0.19	0.18	0.10

Table A1.7 *(continued)*. DoGSite descriptors for the set of cryptic site pairs. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
41	1uk2	apo	115.52	13.18	0.37	0.27	0.69	0.14	0.20	0.21
41	2gz7	holo	148.96	16.88	0.35	0.19	0.67	0.16	0.19	0.12
42	1w50	apo	31.68	18.50	0.50	0.05	0.67	0.19	0.16	0.04
42	3ixj	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
43	1xcg	apo	91.04	15.75	0.41	0.21	0.64	0.15	0.23	0.04
43	1ow3	holo	115.20	19.53	0.40	0.14	0.69	0.16	0.18	0.08
44	1z92	apo	42.88	11.36	0.47	0.20	0.78	0.13	0.09	0.30
44	1py2	holo	46.56	10.13	0.41	0.18	0.78	0.14	0.10	0.14
45	1zah	apo	193.44	21.69	0.31	0.27	0.64	0.19	0.21	0.03
45	2ot1	holo	224.32	19.22	0.34	0.23	0.63	0.20	0.21	0.04
46	2air	apo	228.80	24.65	0.40	0.21	0.65	0.16	0.19	0.03
46	1za1	holo	61.44	11.85	0.37	0.14	0.69	0.21	0.09	0.00
47	2aka	apo	134.08	31.54	0.46	0.14	0.67	0.18	0.18	0.05
47	1yv3	holo	273.76	49.49	0.51	0.12	0.67	0.19	0.16	0.09
48	2ax9	apo	68.80	26.27	0.55	0.08	0.75	0.12	0.15	0.10
48	2piq	holo	99.20	14.98	0.36	0.24	0.76	0.16	0.08	0.02
49	2bf3	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
49	3dhh	holo	71.84	17.32	0.44	0.19	0.69	0.21	0.10	0.01
50	2bls	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
50	3gqz	holo	145.28	25.53	0.43	0.13	0.65	0.18	0.20	0.07
51	2brk	apo	50.88	10.47	0.38	0.28	0.72	0.13	0.18	0.26
51	2gir	holo	78.72	14.91	0.39	0.22	0.70	0.16	0.16	0.20
52	2bu8	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
52	2bu2	holo	33.92	11.69	0.42	0.09	0.82	0.11	0.08	0.32
53	2cga	apo	138.88	27.65	0.45	0.15	0.64	0.19	0.21	0.10
53	1afq	holo	262.08	23.38	0.36	0.19	0.66	0.20	0.18	0.05
54	2cm2	apo	143.20	10.44	0.30	0.33	0.69	0.11	0.24	0.13
54	2h4k	holo	38.24	9.91	0.42	0.12	0.73	0.09	0.21	0.18
55	2gfc	apo	91.84	23.74	0.48	0.08	0.70	0.17	0.14	0.07
55	2jds	holo	353.92	31.28	0.47	0.13	0.69	0.17	0.16	0.11
56	2h4e	apo	33.12	9.10	0.49	0.20	0.68	0.19	0.19	0.00
56	3cfn	holo	65.60	14.20	0.43	0.23	0.69	0.17	0.15	0.00
57	2iyt	apo	101.28	15.00	0.37	0.30	0.66	0.16	0.26	0.03
57	2iyq	holo	146.56	21.65	0.38	0.15	0.67	0.17	0.19	0.05
58	2ohg	apo	161.76	30.75	0.42	0.15	0.67	0.18	0.19	0.05
58	2ohv	holo	73.12	17.21	0.44	0.15	0.66	0.19	0.21	0.09
59	2q8f	apo	0.32	11.74	0.67	0.00	0.74	0.13	0.16	0.11
59	2q8h	holo	2.72	9.57	0.52	0.01	0.77	0.10	0.16	0.16
60	2qfo	apo	100.32	18.77	0.40	0.13	0.69	0.17	0.16	0.07
60	2wi7	holo	172.48	26.32	0.39	0.14	0.71	0.15	0.15	0.16
61	2qlr	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
61	3dc1	holo	57.12	9.08	0.35	0.31	0.73	0.15	0.15	0.15
62	2wgb	apo	61.92	11.27	0.45	0.15	0.72	0.14	0.14	0.08
62	2v57	holo	115.04	14.57	0.41	0.14	0.68	0.18	0.16	0.19
63	2wgq	apo	195.84	56.10	0.62	0.08	0.70	0.16	0.16	0.14
63	1d6y	holo	28.00	23.36	0.67	0.05	0.74	0.15	0.14	0.19
64	2yqc	apo	36.64	16.85	0.59	0.11	0.72	0.15	0.13	0.10
64	2yqs	holo	361.28	33.43	0.41	0.18	0.67	0.18	0.18	0.08

Table A1.7 *(continued)*. DoGSite descriptors for the set of cryptic site pairs. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite[217] (n/a).

Group	PDB-ID	State	Lid	Depth	A/V	Lid/ Hull	Hydro-phob-icity	Acc. Atoms	Don. Atoms	Aro. Atoms
65	2zb1	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
65	2npq	holo	107.52	25.85	0.39	0.13	0.76	0.14	0.12	0.12
66	3b7d	apo	62.40	8.72	0.38	0.31	0.64	0.23	0.18	0.05
66	2al4	holo	50.72	7.55	0.37	0.27	0.66	0.18	0.21	0.05
67	3bl9	apo	216.16	26.68	0.41	0.17	0.66	0.18	0.19	0.12
67	3bl7	holo	240.48	26.63	0.38	0.22	0.69	0.16	0.17	0.21
68	3che	apo	124.80	20.29	0.38	0.18	0.78	0.14	0.12	0.42
68	2iuz	holo	151.36	26.91	0.45	0.12	0.70	0.18	0.18	0.29
69	3dxn	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
69	3hzt	holo	348.96	32.28	0.42	0.17	0.69	0.19	0.16	0.10
70	3fdl	apo	247.52	28.75	0.36	0.24	0.68	0.19	0.16	0.13
70	2yxj	holo	49.60	11.33	0.42	0.13	0.80	0.14	0.07	0.25
71	3gxd	apo	69.92	18.26	0.51	0.10	0.67	0.20	0.15	0.25
71	2wcg	holo	94.56	15.83	0.43	0.17	0.66	0.21	0.16	0.33
72	3h5r	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
72	3h9j	holo	185.92	25.86	0.43	0.13	0.69	0.17	0.17	0.10
73	3hqd	apo	77.12	15.01	0.48	0.14	0.70	0.18	0.16	0.12
73	1q0b	holo	154.72	29.51	0.46	0.11	0.67	0.16	0.19	0.11
74	3kqa	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
74	3lth	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
75	3l7u	apo	238.56	25.34	0.41	0.16	0.68	0.16	0.18	0.14
75	2hvd	holo	134.24	23.63	0.37	0.15	0.64	0.18	0.21	0.15
76	3mn9	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
76	3eks	holo	185.92	22.84	0.29	0.30	0.78	0.14	0.13	0.21
77	3nnu	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
77	3hl7	holo	41.28	15.93	0.43	0.07	0.74	0.15	0.15	0.09
78	3peo	apo	59.84	14.28	0.49	0.13	0.68	0.18	0.18	0.13
78	2bys	holo	39.20	15.45	0.47	0.08	0.72	0.16	0.15	0.21
79	3puw	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
79	1fqc	holo	64.48	22.75	0.43	0.08	0.75	0.12	0.14	0.25
80	4ake	apo	56.80	9.31	0.39	0.25	0.64	0.17	0.21	0.07
80	1ank	holo	333.44	34.40	0.41	0.11	0.64	0.18	0.18	0.06
81	4hb2	apo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
81	4hat	holo	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Table A1.8. Sensitivity towards binding site flexibility for the set of cryptic site pairs. For the residue- and ligand-based analysis, the number of site pairs where the difference between the *RRO* or *RLO* of the ligand-bound structure and the unbound structure exceeds 0.3 or 0.5, respectively.

Method	RRO_{holo}-RRO_{apo} > 0.3	Fraction of the dataset	RLO_{holo}-RLO_{apo} > 0.5	Fraction of the dataset
ConCavity[SURFNET]	35	0.43	23	0.27
PrinCCes	36	0.44	23	0.27
EPOS[BP]	34	0.42	19	0.22
VolSite	29	0.36	21	0.25
PHECOM	28	0.35	25	0.29
fpocket (version 1.0)	33	0.41	30	0.35
PocketPicker	26	0.32	22	0.26
MSPocket	27	0.33	23	0.27

Table A1.8 *(continued)*. Sensitivity towards binding site flexibility for the set of cryptic site pairs. For the residue- and ligand-based analysis, the number of site pairs where the difference between the *RRO* or *RLO* of the ligand-bound structure and the unbound structure exceeds 0.3 or 0.5, respectively.

Method	RRO_{holo}-RRO_{apo} > 0.3	Fraction of the dataset	RLO_{holo}-RLO_{apo} > 0.5	Fraction of the dataset
ConCavity^{LIGSITE}	24	0.30	18	0.21
fpocket (version 2.0)	33	0.41	33	0.39
ghecom	26	0.32	17	0.20
LIGSITE^{CS}	23	0.28	17	0.20
PASS	25	0.31	22	0.26
NanoShaper	39	0.48	33	0.39
Void	31	0.38	20	0.24
PocketAnalyzer^{PCA}	23	0.28	19	0.22
DoGSite	17	0.21	25	0.29
KVFinder	16	0.20	31	0.36
GetCleft	22	0.27	20	0.24
PocketDepth (surface atoms)	28	0.35	23	0.27
CLIPPERS	30	0.37	20	0.24
PocketDepth (polar atoms)	29	0.36	21	0.25
PocketDepth (size)	27	0.33	19	0.22
PocketDepth (non-polar atoms)	35	0.43	23	0.27
CAVITY	9	0.11	12	0.14
trj_cavity	22	0.27	20	0.24
fconv (Delaunay triangulation)	20	0.25	21	0.25
VisGrid	26	0.32	10	0.12
PocketDepth (depth factor)	32	0.40	15	0.18
McVol	22	0.27	10	0.12
MOLE 2.0	28	0.35	29	0.34
SURFNET	11	0.14	10	0.12
VOIDOO	16	0.20	12	0.14
dxTuber	14	0.17	14	0.16
fconv (grid-based)	3	0.04	14	0.16
CAVITATOR	10	0.12	12	0.14
BetaVoid	6	0.07	2	0.02
CAVITY SEARCH	8	0.10	4	0.05
Voronoia	7	0.09	4	0.05
EDTSurf	5	0.06	0	0.00
CAVE	0	0.00	0	0.00
AutoSite	27	0.33	26	0.31
SiteHound (carbon probe)	40	0.49	26	0.31
ConCavity^{PocketFinder}	28	0.35	16	0.19
SiteMap	27	0.33	25	0.29
SiteHound (phosphate probe)	32	0.40	13	0.15
P2RANK	23	0.28	10	0.12
ConCavity^{SURFNET} (conservation)	26	0.32	11	0.13
ConCavity^{PocketFinder} (conservation)	26	0.32	14	0.16
ConCavity^{LIGSITE} (conservation)	17	0.21	12	0.14
MetaPocket 2.0	20	0.25	18	0.21
Binding Response	34	0.42	25	0.29
MolSite	32	0.40	21	0.25

Table A1.8 *(continued)*. Sensitivity towards binding site flexibility for the set of cryptic site pairs. For the residue- and ligand-based analysis, the number of site pairs where the difference between the *RRO* or *RLO* of the ligand-bound structure and the unbound structure exceeds 0.3 or 0.5, respectively.

Method	RRO_{holo}-RRO_{apo} > 0.3	Fraction of the dataset	RLO_{holo}-RLO_{apo} > 0.5	Fraction of the dataset
LISE	31	0.38	18	0.21
G-LoSA	21	0.26	6	0.07
TM-SITE	14	0.17	6	0.07
COFACTOR	23	0.28	14	0.16
S-SITE	2	0.02	5	0.06
COACH	18	0.22	8	0.09
DEPTH (version 2.0)	11	0.14	21	0.25

Table A1.9. The dataset of homology models for *Mycobacterium tuberculosis* proteins. The dataset includes the comparative models from the SWISS-MODEL repository and the corresponding templates (maximum sequence similarity: 40%).

Template (PDB-ID. Chain)	Ligand-Name	Ligand-ID	Ligand-Chain	UniProt-Accession	Model-ID	Sequence Similarity
1knp.A	FAD,SIN	800,801	A	P9WJJ9	59ca81edad01f085524f215d	38%
2d62.A	POP	1293	A	P9WQI5	59c917037aaa0d94ee7022c3	36%
2f7f.A	NIO	601	A	P9WJI7	59c9ff0fa52be7af7a107a41	39%
2j5v.B	PCA,RGP	1368,1369	B	P9WHU9	59c8fca3140550b4cbea4f08	38%
3hfv.A	MTY	416	A	P9WFU3	59c8f3d6b4c268f37f274d9a	39%
3wir.C	BGC,BGC, PO4	801,802, 805	C	P9WN15	59ccfae0a52be722589c2ea5	36%
4cvm.A(1)	ANP	1456	A	P9WJL1	59ca57a9c730d446b09b4838	36%
4cvm.A(2)	UDP,MUB, ALA,FGA	1452,1453, 1454,1455	A	P9WJL1	59ca57a9c730d446b09b4838	36%
5bsh.A	PRO	301	A	P9WHU7	59c9c33f7aaa0db832bebff8	37%
5dir.A	MLE,IIL, SER,ALO, 5BV	201,202, 203,204, 205	A	P9WK99	59cb0ad6573f201ea109ee6a	38%
5ktn.A	SF4,13P	401,402	A	P9WJK1	59c8da70e6291dfcab61acfd	39%

Table A1.10. Correlation between the hitrates and the RMSD between homology models and template structures. The template structures and homology models were aligned in UCSF Chimera[220] (default settings) and the RMSD across all aligned residue pairs was calculated. The Pearson correlation coefficient between the *RRO*- and *RLO*-based top3 hitrates for the model sites and the RMSD of the model-template pairs was calculated. The correlation coefficients are coloured using a colour gradient ranging from green (for a correlation coefficient of 1) to yellow (for a correlation coefficient of 0) to red (for a correlation coefficient of -1).

Method	top3 (*RRO* ≥ 0.5)	top3 (*RLO* ≥ 0.5)
$ConCavity^{SURFNET}$	-0.02	-0.20
PrinCCes	-0.20	-0.29
$EPOS^{BP}$	-0.06	-0.33
VolSite	-0.04	-0.11
PHECOM	-0.36	-0.31
fpocket (version 1.0)	-0.38	-0.64
PocketPicker	0.00	-0.13
MSPocket	-0.15	-0.38
$ConCavity^{LIGSITE}$	-0.28	-0.36
fpocket (version 2.0)	-0.19	-0.41
ghecom	-0.51	-0.49
$LIGSITE^{CS}$	-0.21	-0.22

Table A1.10 *(continued)*. Correlation between the hitrates and the RMSD between homology models and template structures. The template structures and homology models were aligned in UCSF Chimera[220] (default settings) and the RMSD across all aligned rresidue pairs was calculated. The Pearson correlation coefficient between the *RRO*- and *RLO*-based top3 hitrates and the RMSD of the model-template pairs was calculated. The correlation coefficients are coloured using a colour gradient ranging from green (for a correlation coefficient of 1) to yellow (for a correlation coefficient of 0) to red (for a correlation coefficient of -1).

Method	top3 (*RRO* ≥ 0.5)	top3 (*RLO* ≥ 0.5)
PASS	-0.72	-0.74
NanoShaper	0.56	0.50
Void	-0.28	-0.18
PocketAnalyzer^{PCA}	-0.27	-0.32
DoGSite	-0.44	-0.33
KVFinder	-0.04	-0.17
GetCleft	-0.54	-0.47
PocketDepth (surface atom count)	-0.01	-0.21
CLIPPERS	-0.13	-0.24
PocketDepth (polar atom count)	0.04	-0.15
PocketDepth (size)	0.37	-0.07
PocketDepth (non-polar atom count)	0.11	-0.08
CAVITY	-0.56	-0.55
trj_cavity	-0.18	-0.25
fconv (Delaunay triangulation)	-0.50	-0.48
VisGrid	-0.26	-0.34
PocketDepth (depth factor)	0.09	-0.13
McVol	0.13	-0.03
MOLE 2.0	0.11	0.10
SURFNET	-0.11	-0.46
VOIDOO	-0.42	-0.37
dxTuber	-0.04	-0.50
fconv (grid-based)	-0.35	-0.32
CAVITATOR	-0.06	-0.07
BetaVoid	-0.25	-0.43
CAVITY SEARCH	-0.31	-0.48
Voronoia	-0.24	-0.45
EDTSurf	-0.46	-0.58
CAVE	n/a	-0.07
AutoSite	-0.19	-0.26
SiteHound (carbon probe)	0.17	-0.09
ConCavity^{PocketFinder}	-0.15	-0.09
SiteMap	-0.39	-0.53
SiteHound (phosphate probe)	-0.25	-0.22
P2RANK	-0.34	-0.35
ConCavity^{SURFNET} (conservation)	-0.02	-0.20
ConCavity^{PocketFinder} (conservation)	0.03	-0.01
MetaPocket 2.0	-0.34	-0.43
ConCavity^{LIGSITE} (conservation)	-0.45	-0.49
MolSite	-0.15	-0.19
Binding Response	-0.06	-0.09
LISE	0.01	-0.03
G-LoSA	0.26	0.15
S-SITE	-0.11	-0.25
TM-SITE	0.33	0.15
COFACTOR	0.17	-0.01
COACH	0.35	0.11
DEPTH (version 2.0)	-0.34	-0.44

Table A1.11. The dataset of allosteric binding sites. The proteins and ligands of the diverse ASBench dataset[221] together with the correspoding sites' DoGSite descriptors are summarised below. If several binding sites for one structure were used, the index of the binding site (#) is given. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite (n/a).

PDB-ID	Chain- ID	Ligand-Name	Ligand-ID	#	Lid	Depth	A/V	Lid/ Hull	Acc. Atoms	Don. Atoms	Aro. Atoms
11bg	A	U2G	131	1	91.36	13.42	0.44	0.27	0.17	0.20	0.08
1b86	D	DG2	701	1	55.84	9.26	0.36	0.28	0.08	0.22	0.00
1ce8	A	ORN	5011	1	31.20	16.65	0.56	0.12	0.20	0.15	0.11
1ce8	A	IMP	5012	2	45.28	12.63	0.44	0.22	0.17	0.25	0.05
1dku	A	AP2	1002	1	236.32	32.24	0.42	0.20	0.16	0.20	0.07
1efa	A	NPF	901	1	130.56	23.77	0.41	0.10	0.16	0.17	0.14
1em6	A	CP4	862	1	47.68	19.58	0.61	0.13	0.19	0.16	0.14
1f2u	A	ATP	901	1	283.84	39.27	0.44	0.11	0.18	0.19	0.12
1fiy	A	ASP	884	1	46.72	15.44	0.52	0.11	0.20	0.17	0.07
1fx2	A	DTT	1124	1	0.64	8.50	0.64	0.00	0.12	0.20	0.19
1gz3	A	FUM	605	1	174.72	35.68	0.57	0.09	0.16	0.15	0.11
1h5s	A	TMP	1292	1	148.80	28.42	0.53	0.10	0.17	0.16	0.11
1h78	A	DCP	1589	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
1h9g	A	COA; MYR	1228; 1229	1	227.52	27.36	0.46	0.11	0.15	0.16	0.18
1hkb	A	G6P	919	1	58.72	31.01	0.57	0.15	0.23	0.22	0.01
1i7s	A	TRP	601	1	59.84	21.16	0.47	0.11	0.18	0.16	0.11
1jlr	A	GTP	303	1	70.40	12.77	0.43	0.19	0.13	0.19	0.14
1kz8	A	PFE	738	1	62.72	8.85	0.36	0.26	0.16	0.14	0.00
1m8p	A	PPS	574	1	127.20	27.15	0.46	0.16	0.12	0.19	0.11
1nh8	A	HIS	289	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
1njj	A	GET	601	1	118.40	19.86	0.39	0.14	0.18	0.15	0.06
1nsg	A	RAD	108	1	165.60	14.78	0.29	0.28	0.09	0.11	0.53
1of6	A	DTY	1370	1	101.76	35.43	0.55	0.04	0.17	0.18	0.09
1pcq	A	ADP; AF3	600; 602	1	133.28	32.86	0.48	0.09	0.19	0.16	0.02
1pfk	A	ADP	326	1	42.56	22.07	0.48	0.09	0.15	0.24	0.09
1qp0	A	HPA	599	1	141.60	30.76	0.49	0.09	0.17	0.18	0.12
1qw7	A	EBP	702	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
1shj	A	NXN	401	1	36.16	21.78	0.47	0.04	0.14	0.15	0.16
1t49	A	892	301	1	64.16	12.28	0.41	0.19	0.15	0.08	0.20
1uxv	A	AMP	1503	1	103.04	17.96	0.39	0.23	0.13	0.22	0.09
1w0f	A	STR	1499	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
1w96	A	S1A	1567	1	179.36	25.66	0.41	0.20	0.15	0.19	0.25
1xjf	A	DTP	1002	1	9.76	13.35	0.61	0.05	0.17	0.18	0.04
1xtu	A; D; E; H	CTP; CTP; CTP; CTP	1260; 4260; 5260; 8260	1	41.92	23.20	0.54	0.07	0.16	0.11	0.04
1z8d	A	AMP	900	1	26.40	11.26	0.52	0.17	0.15	0.26	0.37
1z8d	A	ADE	902	2	22.08	12.91	0.63	0.09	0.13	0.14	0.24
1zds	A	ACM	2500	1	80.64	26.71	0.51	0.09	0.16	0.17	0.18
2a69	C	RPT	8001	1	26.88	16.99	0.63	0.20	0.19	0.17	0.05
2be9	B	CTP	401	1	55.04	9.47	0.38	0.25	0.14	0.08	0.00

Table A1.11 *(continued)*. The dataset of allosteric binding sites. The proteins and ligands of the diverse ASBench dataset[221] together with the correspoding sites' DoGSite descriptors are summarised below. If several binding sites for one structure were used, the index of the binding site (#) is given. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite (n/a).

PDB-ID	Chain- ID	Ligand-Name	Ligand-ID	#	Lid	Depth	A/V	Lid/ Hull	Acc. Atoms	Don. Atoms	Aro. Atoms
2bu6	A	TF2	1386	1	19.68	10.58	0.45	0.06	0.11	0.07	0.31
2bu7	A	TF3	1386	1	49.28	17.41	0.44	0.08	0.17	0.14	0.15
2bu8	A	TF4	1379	1	5.44	11.24	0.56	0.02	0.12	0.19	0.14
2bxa	A	C1F	2001	1	172.96	26.84	0.42	0.17	0.16	0.15	0.08
2d5x	B	L35	1201	1	87.84	19.58	0.42	0.17	0.20	0.19	0.12
2him	A	ASN	8001	1	14.24	20.92	0.57	0.01	0.16	0.17	0.08
2hvw	A	DCP	1201	1	80.96	10.53	0.29	0.40	0.23	0.29	0.17
2i7n	A	ACO	796	1	150.08	26.70	0.48	0.12	0.16	0.19	0.11
2jfn	A	UMA	1286	1	145.12	25.79	0.41	0.21	0.20	0.14	0.00
2jhr	A	PBQ	1780	1	228.48	50.45	0.54	0.13	0.18	0.17	0.08
2ldb	C	FBP	6	1	44.96	21.96	0.57	0.11	0.12	0.16	0.14
2puv	B	UD1	5003	1	25.92	9.31	0.48	0.16	0.22	0.26	0.11
2q5o	A	PPY	5003	1	47.84	26.40	0.51	0.07	0.15	0.20	0.06
2q8m	A	BG6	340	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
2rd5	A	ARG	1000	1	101.28	20.99	0.41	0.16	0.24	0.13	0.07
2rde	A	C2E	301	1	45.12	11.98	0.41	0.12	0.17	0.28	0.00
2v4y	A	GTP	1242	1	131.20	24.83	0.39	0.25	0.16	0.24	0.08
2vd3	A	HIS	1290	1	45.12	7.78	0.41	0.26	0.16	0.22	0.00
2vpr	A	TDC	1206	1	251.84	30.30	0.43	0.13	0.16	0.15	0.13
2w4i	B	VGA	1256	1	154.56	33.45	0.49	0.08	0.17	0.18	0.09
2wu1	A	16G	1267	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
2xjc	A	B4P	1490	1	0.80	11.42	0.64	0.00	0.17	0.15	0.03
2y0p	A	ACO	2228	1	88.80	31.24	0.57	0.17	0.18	0.18	0.04
2yhd	A	AV6	1921	1	46.40	9.62	0.46	0.18	0.12	0.08	0.08
2ylo	A	YLO	1922	1	56.16	14.06	0.45	0.19	0.16	0.09	0.23
2zmf	A	CMP	2001	1	41.12	17.39	0.47	0.06	0.14	0.14	0.20
3bz7	A	BL4	800	1	286.40	49.32	0.50	0.13	0.20	0.18	0.07
3csm	A	TRP	300	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
3dc2	A	SER	600	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
3e5u	A	3C4	501	1	125.12	32.25	0.50	0.08	0.16	0.14	0.11
3eps	A	AMP	1604	1	38.56	17.19	0.43	0.09	0.16	0.17	0.17
3ete	A	GTP	553	1	64.96	24.00	0.51	0.10	0.18	0.19	0.01
3ete	A; B; C; C; D; F	H3P; H3P; H3P; H3P; H3P; H3P	552; 552; 552; 554; 552; 552	2	25.92	11.51	0.40	0.46	0.16	0.12	0.09
3etg	A; E	GWD; GWD	552; 552	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
3f6g	B	ILE	1	1	72.32	32.09	0.53	0.06	0.17	0.16	0.09
3f9n	A	38M	324	1	36.96	8.10	0.44	0.15	0.17	0.06	0.17
3g86	A	T18	571	1	87.36	21.50	0.51	0.10	0.16	0.14	0.11
3gcd	A	IHP	214	1	174.40	18.88	0.40	0.18	0.18	0.24	0.08

Table A1.11 *(continued)*. The dataset of allosteric binding sites. The proteins and ligands of the diverse ASBench dataset[221] together with the correspoding sites' DoGSite descriptors are summarised below. If several binding sites for one structure were used, the index of the binding site (#) is given. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite (n/a).

PDB-ID	Chain- ID	Ligand-Name	Ligand-ID	#	Lid	Depth	A/V	Lid/ Hull	Acc. Atoms	Don. Atoms	Aro. Atoms
3h1v	X	TK1	501	1	168.64	21.11	0.34	0.19	0.17	0.14	0.09
3h30	A	RFZ	337	1	26.72	9.41	0.47	0.11	0.16	0.12	0.06
3h6o	A	FBP	541	1	270.40	43.31	0.50	0.11	0.16	0.18	0.07
3hnc	A	TTP	802	1	16.80	13.04	0.57	0.10	0.17	0.22	0.06
3ho6	A	IHP	270	1	62.24	14.74	0.51	0.07	0.18	0.19	0.07
3ho8	A	COA	2001	1	19.84	9.95	0.46	0.39	0.23	0.19	0.00
3ifa	A	AMP	339	1	149.76	25.06	0.47	0.12	0.20	0.20	0.07
3irh	A	DGT	458	1	45.44	16.75	0.54	0.22	0.13	0.24	0.30
3kcg	I	NTP	433	1	45.44	7.27	0.39	0.30	0.11	0.19	0.00
3kf0	A	4DX	201	1	86.56	13.98	0.42	0.25	0.12	0.14	0.00
3kgf	A	PHE	9003	1	33.28	25.33	0.49	0.06	0.17	0.16	0.10
3kgf	A	TRP	9004	2	42.08	15.30	0.48	0.16	0.15	0.13	0.08
3l3v	A	SUC	1	1	174.72	22.99	0.45	0.11	0.16	0.15	0.11
3laj	A	ARG	200	1	195.84	45.37	0.56	0.09	0.19	0.21	0.04
3lsw	A	4MP	801	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
3lu6	A	IMX	587	1	73.92	19.28	0.38	0.16	0.14	0.09	0.07
3lw0	A	CCX	1	1	46.88	16.25	0.46	0.07	0.16	0.12	0.14
3m6f	A	BJZ	1	1	58.40	11.29	0.39	0.12	0.09	0.08	0.21
3mks	D	C1C	1	1	55.20	12.87	0.42	0.14	0.19	0.20	0.12
3mw9	A	NAI	604	1	64.80	27.34	0.56	0.13	0.20	0.20	0.17
3mwb	B	PHE	311	1	1.12	11.52	0.59	0.00	0.14	0.14	0.12
3n1v	F	3N1	1	1	80.00	18.51	0.45	0.11	0.15	0.18	0.19
3o2m	A	46A	701	1	46.72	28.29	0.64	0.08	0.17	0.20	0.10
3o96	A	IQO	444	1	402.88	40.53	0.39	0.13	0.16	0.19	0.12
3pg9	A	TYR	339	1	133.76	38.80	0.55	0.11	0.17	0.20	0.06
3pjg	A	UGA	902	1	116.32	16.08	0.38	0.18	0.15	0.17	0.08
3pma	B	SCR	248	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a
3ptz	A	UDX	501	1	179.68	39.55	0.53	0.10	0.16	0.18	0.10
3pxf	A	2AN; 2AN	304; 305	1	131.36	27.39	0.39	0.09	0.14	0.14	0.07
3pyy	A	3YY	538	1	38.88	18.02	0.47	0.09	0.14	0.07	0.13
3qel	B	QEL	1	1	181.28	32.28	0.50	0.11	0.19	0.15	0.10
3qh0	A	PLM	625	1	272.16	45.61	0.48	0.09	0.17	0.16	0.11
3qop	A	CMP	301	1	120.64	25.44	0.44	0.11	0.17	0.17	0.06
3r1r	A	ATP	762	1	66.24	12.67	0.43	0.18	0.11	0.17	0.10
3uo9	B	04A	2	1	24.80	18.52	0.58	0.14	0.13	0.16	0.26
3vq8	A	BCU	1006	1	66.40	12.96	0.41	0.16	0.17	0.10	0.15
3zcw	A	4A2	1366	1	123.04	25.90	0.40	0.11	0.17	0.16	0.13
4ahs	A	AKH	1216	1	57.44	12.69	0.41	0.13	0.16	0.15	0.16
4aw0	A	MJF	600	1	72.32	13.32	0.42	0.18	0.13	0.17	0.07
4b1f	A	KRH	1256	1	151.04	19.53	0.35	0.19	0.19	0.12	0.20
4b2d	A	SER	1532	1	68.16	21.05	0.46	0.09	0.15	0.19	0.15
4bbg	A	V02	1370	1	239.20	31.06	0.43	0.13	0.18	0.19	0.09

Table A1.11 *(continued)*. The dataset of allosteric binding sites. The proteins and ligands of the diverse ASBench dataset[221] together with the correspoding sites' DoGSite descriptors are summarised below. If several binding sites for one structure were used, the index of the binding site (#) is given. The number of acceptor (acc.), donor (don.), and aromatic (aro.) atoms was normalised by the number of site atoms. Descriptors could not be calculated for sites in the proximity of the ligand which were not detected by DoGSite (n/a).

PDB-ID	Chain- ID	Ligand-Name	Ligand-ID	#	Lid	Depth	A/V	Lid/ Hull	Acc. Atoms	Don. Atoms	Aro. Atoms
4bo2	A	36K	1248	1	89.92	19.81	0.46	0.09	0.17	0.14	0.11
4bqh	A	9VU	1539	1	39.36	13.33	0.43	0.08	0.19	0.16	0.11
4bzb	A; C; D	DGT; DGT	800; 900	1	184.64	55.70	0.59	0.10	0.17	0.20	0.11
4c7b	A	BVB	1395	1	322.08	27.52	0.33	0.22	0.17	0.18	0.13
4cfe	A	992	1553	1	140.80	31.90	0.45	0.16	0.17	0.19	0.10
4e6c	A	0O8	500	1	56.00	16.17	0.42	0.10	0.13	0.11	0.11
4eaj	C	AMP	403	1	435.36	34.60	0.41	0.16	0.14	0.16	0.10
4ebw	A	0PF	700	1	222.08	30.74	0.42	0.10	0.17	0.15	0.11
4ej8	A	1F1	106	1	45.92	10.03	0.46	0.25	0.13	0.10	0.08
4eo6	A	0S2	600	1	139.68	22.56	0.38	0.28	0.14	0.18	0.15
4etz	A	C2E	501	1	69.92	14.24	0.42	0.14	0.15	0.22	0.04
4g1n	A	NZT	603	1	191.84	51.24	0.57	0.10	0.18	0.14	0.10
4g8o	D	96P	403	1	28.32	11.36	0.46	0.08	0.10	0.15	0.22
4hsg	A	KTD	601	1	18.08	9.78	0.48	0.05	0.16	0.14	0.04
4i1r	A	LZU	801	1	83.20	12.96	0.40	0.12	0.13	0.14	0.10
4ig3	A	J94	609	1	148.32	34.07	0.53	0.13	0.14	0.14	0.11
4io8	A	3FD	401	1	274.40	29.25	0.42	0.12	0.20	0.19	0.11
4ja8	A	1K9	502	1	5.92	18.06	0.61	0.01	0.13	0.10	0.18
4jaf	A	NAI	503	1	25.76	8.66	0.48	0.19	0.18	0.20	0.20
4kfb	A	1QP	601	1	155.36	41.48	0.54	0.13	0.14	0.14	0.13
4kko	A	1RE	601	1	21.92	13.49	0.53	0.09	0.12	0.10	0.26
4mbs	A	MRV	1101	1	340.48	25.92	0.36	0.22	0.16	0.12	0.29
4mdk	A	U94	201	1	295.68	28.03	0.42	0.14	0.15	0.14	0.18
4mra	A	QUE	901	1	54.88	15.93	0.44	0.12	0.19	0.15	0.03
4nbn	B	2J7	401	1	139.04	38.43	0.52	0.10	0.19	0.20	0.08
4nld	A	2N7	601	1	39.36	9.70	0.40	0.20	0.13	0.13	0.23
4oo9	A	2U8	4005	1	123.52	45.69	0.54	0.06	0.17	0.14	0.13
4p3h	A	25G	201	1	83.52	11.11	0.36	0.23	0.11	0.11	0.20

REFERENCES

[1] Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288.

[2] Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849–857.

[3] David, C. W. Voronoi polyhedra as structure probes in large molecular systems. *Biopolymers* **1988**, *27* (2), 339–344.

[4] Voorintholt, R.; Kosters, M. T.; Vegter, G.; Vriend, G.; Hol, W.G.J. A very fast program for visualizing protein surfaces, channels and cavities. *J. Mol. Graphics* **1989**, *7* (4), 243–245.

[5] Lewis, R. A. Determination of clefts in receptor structures. *J. Comput.-Aided Mol. Des.* **1989**, *3* (2), 133–147.

[6] Ho, C. M.; Marshall, G. R. Cavity Search: An algorithm for the isolation and display of cavity-like binding regions. *J. Comput.-Aided Mol. Des.* **1990**, *4* (4), 337–354.

[7] Levitt, D. G.; Banaszak, L. J. POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graphics* **1992**, *10* (4), 229–234.

[8] Delaney, J. S. Finding and filling protein cavities using cellular logic operations. *J. Mol. Graphics* **1992**, *10* (3), 174-7, 163.

[9] Del Carpio, C. A.; Takahashi, Y.; Sasaki, S. A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I) Search for pocket regions. *J. Mol. Graphics* **1993**, *11* (1), 23-9, 42.

[10] Williams, M. A.; Goodfellow, J. M.; Thornton, J. M. Buried waters and internal cavities in monomeric proteins. *Protein Sci.* **1994**, *3* (8), 1224–1235.

[11] Kleywegt, G. J.; Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *50* (Pt 2), 178–185.

[12] Laskowski, R. A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics* **1995**, *13* (5), 323-30, 307-8.

[13] Casari, G.; Sander, C.; Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Mol. Biol.* **1995**, *2* (2), 171–178.

[14] Masuya, M.; Doi, J. Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. *J. Mol. Graphics* **1995**, *13* (6), 331–336.

[15] Peters, K. P.; Fauck, J.; Frömmel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **1996**, *256* (1), 201–213.

[16] Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38* (3), 305–320.

[17] Lichtarge, O.; Bourne, H. R.; Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **1996**, *257* (2), 342–358.

[18] Ruppert, J.; Welch, W.; Jain, A. N. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* **1997**, *6* (3), 524–533.

[19] Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15* (6), 359.

[20] Exner, T.; Keil, M.; Moeckel, G.; Brickmann, J. Identifification of substrate channels and protein cavities. *J. Mol. Model.* **1998**, *4* (10), 340–343.

[21] Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7* (9), 1884–1897.

[22] Bliznyuk, A. A.; Gready, J. E. Simple method for locating possible ligand binding sites on protein surfaces. *J. Comput. Chem.* **1999**, *20* (9), 983–988.

[23] Brady, G. P.; Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14* (4), 383–401.
[24] Stahl, M.; Taroni, C.; Schneider, G. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng., Des. Sel.* **2000**, *13* (2), 83–88.
[25] Elcock, A. H. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* **2001**, *312* (4), 885–896.
[26] Landgraf, R.; Xenarios, I.; Eisenberg, D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **2001**, *307* (5), 1487–1502.
[27] Aloy, P.; Querol, E.; Aviles, F. X.; Sternberg, M. J. E. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **2001**, *311* (2), 395–408.
[28] Chakravarty, S.; Bhinge, A.; Varadarajan, R. A procedure for detection and quantitation of cavity volumes proteins. Application to measure the strength of the hydrophobic driving force in protein folding. *J. Biol. Chem.* **2002**, *277* (35), 31345–31353.
[29] Pupko, T.; Bell, R. E.; Mayrose, I.; Glaser, F.; Ben-Tal, N. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **2002**, *18 Suppl 1*, 7.
[30] Binkowski, T. A.; Adamian, L.; Liang, J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.* **2003**, *332* (2), 505–526.
[31] Binkowski, T. A. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.* **2003**, *31* (13), 3352–3355.
[32] Glaser, F.; Pupko, T.; Paz, I.; Bell, R. E.; Bechor-Shental, D.; Martz, E.; Ben-Tal, N. ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **2003**, *19* (1), 163–164.
[33] Stark, A. Annotation in three dimensions. PINTS: Patterns In Non-homologous Tertiary Structures. *Nucleic Acids Res.* **2003**, *31* (13), 3341–3344.
[34] Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52* (2), 137–145.
[35] del Sol Mesa, A.; Pazos, F.; Valencia, A. Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **2003**, *326* (4), 1289–1302.
[36] Liang, M. P. WebFEATURE: An interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Res.* **2003**, *31* (13), 3324–3327.
[37] Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21* (4), 289–307.
[38] Binkowski, T. A.; Freeman, P.; Liang, J. pvSOAR: Detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Res.* **2004**, *32* (Web Server issue), W555-8.
[39] Chelliah, V.; Chen, L.; Blundell, T. L.; Lovell, S. C. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **2004**, *342* (5), 1487–1504.
[40] Bhinge, A.; Chakrabarti, P.; Uthanumallian, K.; Bajaj, K.; Chakraborty, K.; Varadarajan, R. Accurate detection of protein:ligand binding sites using molecular dynamics simulations. *Structure* **2004**, *12* (11), 1989–1999.
[41] An, J.; Totrov, M.; Abagyan, R. Comprehensive identification of “druggable” protein ligand binding sites. *Genome Inform.* **2004**, *15* (2), 31–41.
[42] Panchenko, A. R.; Kondrashov, F.; Bryant, S. H. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.* **2004**, *13* (4), 884–892.
[43] Innis, C. A.; Anand, A. P.; Sowdhamini, R. Prediction of functional sites in proteins using conserved functional group analysis. *J. Mol. Biol.* **2004**, *337* (4), 1053–1068.

[44] Bate, P.; Warwicker, J. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J. Mol. Biol.* **2004**, *340* (2), 263–276.
[45] Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. Network analysis of protein structures identifies functional residues. *J. Mol. Biol.* **2004**, *344* (4), 1135–1146.
[46] Laskowski, R. A.; Watson, J. D.; Thornton, J. M. ProFunc: A server for predicting protein function from 3D structure. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W89-93.
[47] Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005**, *21* (9), 1908–1916.
[48] An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics* **2005**, *4* (6), 752–761.
[49] Nimrod, G.; Glaser, F.; Steinberg, D. M.; Ben-Tal, N.; Pupko, T. *In silico* identification of functional regions in proteins. *Bioinformatics* **2005**, *21 Suppl 1*, i328-37.
[50] Chang, D. T.-H.; Oyang, Y.-J.; Lin, J.-H. MEDock: A web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. *Nucleic Acids Res.* **2005**, *33* (Web Server), W233-W238.
[51] Ben-Shimon, A.; Eisenstein, M. Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. *J. Mol. Biol.* **2005**, *351* (2), 309–326.
[52] Liang, S.; Zhang, C.; Liu, S.; Zhou, Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.* **2006**, *34* (13), 3698–3707.
[53] Huang, B.; Schroeder, M. LIGSITE^{CSC}: Predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19.
[54] Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins* **2006**, *63* (4), 892–906.
[55] Pazos, F.; Rausell, A.; Valencia, A. Phylogeny-independent detection of functional residues. *Bioinformatics* **2006**, *22* (12), 1440–1448.
[56] Coleman, R. G.; Sharp, K. A. Travel depth, a new shape descriptor for macromolecules: Application to ligand binding. *J. Mol. Biol.* **2006**, *362* (3), 441–458.
[57] Ming, D.; Wall, M. E. Interactions in native binding sites cause a large change in protein dynamics. *J. Mol. Biol.* **2006**, *358* (1), 213–223.
[58] Petrek, M.; Otyepka, M.; Banáš, P.; Kosinová, P.; Koča, J.; Damborský, J. CAVER: A new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinf.* **2006**, *7*, 316.
[59] Kinoshita, K.; Murakami, Y.; Nakamura, H. eF-seek: Prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W398-402.
[60] López, G.; Valencia, A.; Tress, M. L. firestar - Prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W573-7.
[61] Wei, Y.; Ko, J.; Murga, L. F.; Ondrechen, M. J. Selective prediction of interaction sites in protein structures with THEMATICS. *BMC Bioinf.* **2007**, *8*, 119.
[62] Zhang, X.; Bajaj, C. Extraction, quantification and visualization of protein pockets. *Comput. Syst. Bioinformatics Conf.* **2007**, *6*, 275–286.
[63] Goyal, K.; Mohanty, D.; Mande, S. C. PAR-3D: A server to predict protein active site residues. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W503-5.
[64] Pettit, F. K.; Bare, E.; Tsai, A.; Bowie, J. U. HotPatch: A statistical approach to finding biologically relevant features on protein surfaces. *J. Mol. Biol.* **2007**, *369* (3), 863–879.
[65] Halgren, T. A. New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* **2007**, *69* (2), 146–148.

[66] Brylinski, M.; Prymula, K.; Jurkowski, W.; Kochańczyk, M.; Stawowczyk, E.; Konieczny, L.; Roterman, I. Prediction of functional sites based on the fuzzy oil drop model. *PLoS Comput. Biol.* **2007**, *3* (5), e94.
[67] Innis, C. A. siteFiNDER|3D: A web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W489-94.
[68] Eyrisch, S.; Helms, V. Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.* **2007**, *50* (15), 3457–3464.
[69] Kawabata, T.; Go, N. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins* **2007**, *68* (2), 516–529.
[70] Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1*, 7.
[71] Zhong, S.; MacKerell, A. D. Binding Response: A descriptor for selecting ligand binding site on protein surfaces. *J. Chem. Inf. Model.* **2007**, *47* (6), 2303–2315.
[72] Xie, L.; Bourne, P. E. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinf.* **2007**, *8 Suppl 4*, S9.
[73] Petrek, M.; Kosinová, P.; Koča, J.; Otyepka, M. MOLE: A Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure* **2007**, *15* (11), 1357–1363.
[74] Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.* **2007**, *47* (2), 400–406.
[75] Tseng, Y. Y.; Liang, J. Predicting enzyme functional surfaces and locating key residues automatically from structures. *Ann. Biomed. Eng.* **2007**, *35* (6), 1037–1042.
[76] Li, B.; Turuvekere, S.; Agrawal, M.; La, D.; Ramani, K.; Kihara, D. Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins* **2008**, *71* (2), 670–683.
[77] Brylinski, M.; Skolnick, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (1), 129–134.
[78] Kalidas, Y.; Chandra, N. PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins. *J. Struct. Biol.* **2008**, *161* (1), 31–42.
[79] Sankararaman, S.; Sjölander, K. INTREPID - INformation-theoretic TREe traversal for Protein functional site IDentification. *Bioinformatics* **2008**, *24* (21), 2445–2452.
[80] Morita, M.; Nakamura, S.; Shimizu, K. Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. *Proteins* **2008**, *73* (2), 468–479.
[81] Ye, K.; Feenstra, K. A.; Heringa, J.; Ijzerman, A. P.; Marchiori, E. Multi-RELIEF: A method to recognize specificity determining residues from multiple sequence alignments using a machine-learning approach for feature weighting. *Bioinformatics* **2008**, *24* (1), 18–25.
[82] Ming, D.; Cohn, J. D.; Wall, M. E. Fast dynamics perturbation analysis for prediction of protein functional sites. *BMC Struct. Biol.* **2008**, *8*, 5.
[83] Nimrod, G.; Schushan, M.; Steinberg, D. M.; Ben-Tal, N. Detection of functionally important regions in “hypothetical proteins” of known structure. *Structure* **2008**, *16* (12), 1755–1763.
[84] Bock, M. E.; Garutti, C.; Guerra, C. Cavity detection and matching for binding site recognition. *Theor. Comput. Sci.* **2008**, *408* (2-3), 151–162.
[85] Bordner, A. J. Predicting small ligand binding sites in proteins using backbone structure. *Bioinformatics* **2008**, *24* (24), 2865–2871.
[86] Fischer, J. D.; Mayer, C. E.; Söding, J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* **2008**, *24* (5), 613–620.
[87] Ho, B. K.; Gruswitz, F. HOLLOW: Generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct. Biol.* **2008**, *8*, 49.
[88] Kim, D.; Cho, C.-H.; Cho, Y.; Ryu, J.; Bhak, J.; Kim, D.-S. Pocket extraction on proteins via the Voronoi diagram of spheres. *J. Mol. Graphics Modell.* **2008**, *26* (7), 1104–1112.

[89] Huang, B. MetaPocket: A meta approach to improve protein ligand binding site prediction. *OMICS* **2009**, *13* (4), 325–330.
[90] Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D. R.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-based identification of druggable "hot spots" of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009**, *25* (5), 621–627.
[91] Xu, D.; Zhang, Y. Generating triangulated macromolecular surfaces by Euclidean Distance Transform. *PLoS ONE* **2009**, *4* (12), e8140.
[92] Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.
[93] Rother, K.; Hildebrand, P. W.; Goede, A.; Gruening, B.; Preissner, R. Voronoia: Analyzing packing in protein structures. *Nucleic Acids Res.* **2009**, *37* (Database issue), 5.
[94] Ghersi, D.; Sanchez, R. EasyMIFS and SiteHound: A toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **2009**, *25* (23), 3185–3186.
[95] Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.* **2009**, *5* (12), e1000585.
[96] Petrey, D.; Fischer, M.; Honig, B. Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (41), 17377–17382.
[97] Comin, M.; Guerra, C.; Dellaert, F. Binding balls: Fast detection of binding sites using a property of spherical Fourier transform. *J. Comput. Biol.* **2009**, *16* (11), 1577–1591.
[98] Liu, B.; Wang, X.; Lin, L.; Tang, B.; Dong, Q.; Wang, X. Prediction of protein binding sites in protein structures using hidden Markov support vector machine. *BMC Bioinf.* **2009**, *10*, 381.
[99] Talavera, D.; Laskowski, R. A.; Thornton, J. M. WSsas: A web service for the annotation of functional residues through structural homologues. *Bioinformatics* **2009**, *25* (9), 1192–1194.
[100] Tong, W.; Wei, Y.; Murga, L. F.; Ondrechen, M. J.; Williams, R. J. Partial Order Optimum Likelihood (POOL): Maximum likelihood prediction of protein active site residues using 3D Structure and sequence properties. *PLoS Comput. Biol.* **2009**, *5* (1), e1000266.
[101] Tseng, Y. Y.; Dupree, C.; Chen, Z. J.; Li, W.-H. SplitPocket: Identification of protein functional surfaces and characterization of their spatial patterns. *Nucleic Acids Res.* **2009**, *37* (Web Server issue), W384-9.
[102] Tseng, Y. Y.; Chen, Z. J.; Li, W.-H. fPOP: Footprinting functional pockets of proteins by comparative spatial patterns. *Nucleic Acids Res.* **2010**, *38* (Database issue), D288-95.
[103] Wass, M. N.; Kelley, L. A.; Sternberg, M. J. E. 3DLigandSite: Predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W469-73.
[104] Konc, J.; Janežič, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26* (9), 1160–1168.
[105] Mehio, W.; Kemp, G. J. L.; Taylor, P.; Walkinshaw, M. D. Identification of protein binding surfaces using surface triplet propensities. *Bioinformatics* **2010**, *26* (20), 2549–2555.
[106] Yu, J.; Zhou, Y.; Tanaka, I.; Yao, M. Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics* **2010**, *26* (1), 46–52.
[107] Sankararaman, S.; Sha, F.; Kirsch, J. F.; Jordan, M. I.; Sjölander, K. Active site prediction using evolutionary and structural information. *Bioinformatics* **2010**, *26* (5), 617–624.
[108] Tripathi, A.; Kellogg, G. E. A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins* **2010**, *78* (4), 825–842.
[109] Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.* **2010**, *50* (11), 2041–2052.
[110] Kawabata, T. Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins* **2010**, *78* (5), 1195–1211.

[111] Till, M. S.; Ullmann, G. M. McVol - A program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. *J. Mol. Model.* **2010**, *16* (3), 419–429.
[112] Schmidtke, P.; Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **2010**, *53* (15), 5858–5867.
[113] Andersson, C. D.; Chen, B. Y.; Linusson, A. Mapping of ligand-binding cavities in proteins. *Proteins* **2010**, *78* (6), 1408–1422.
[114] Thangudu, R. R.; Tyagi, M.; Shoemaker, B. A.; Bryant, S. H.; Panchenko, A. R.; Madej, T. Knowledge-based annotation of small molecule binding sites in proteins. *BMC Bioinf.* **2010**, *11*, 365.
[115] Buša, J.; Hayryan, S.; Hu, C.-K.; Skřivánek, J.; Wu, M.-C. CAVE: A package for detection and quantitative analysis of internal cavities in a system of overlapping balls: Application to proteins. *Comput. Phys. Commun.* **2010**, *181* (12), 2116–2125.
[116] Kasahara, K.; Kinoshita, K.; Takagi, T. Ligand-binding site prediction of proteins based on known fragment-fragment interactions. *Bioinformatics* **2010**, *26* (12), 1493–1499.
[117] Voss, N. R.; Gerstein, M. 3V: Cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W555-62.
[118] Phillips, M.; Georgiev, I.; Dehof, A. K.; Nickels, S.; Marsalek, L.; Lenhof, H.-P.; Hildebrandt, A.; Slusallek, P. Measuring properties of molecular surfaces using ray casting. In *2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW)*; IEEE, 2010 - 2010; pp 1–7.
[119] Bajaj, C.; Gillette, A.; Goswami, S.; Kwon, B. J.; Rivera, J. Complementary space for enhanced uncertainty and dynamics visualization. In *Topological Methods in Data Analysis and Visualization*; Pascucci, V., Tricoche, X., Hagen, H., Tierny, J., Eds.; Mathematics and Visualization; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; pp 217–228.
[120] Singh, T.; Biswas, D.; Jayaram, B. AADS - An automated active site identification, docking, and scoring protocol for protein targets based on physicochemical descriptors. *J. Chem. Inf. Model.* **2011**, *51* (10), 2515–2527.
[121] Mitternacht, S.; Berezovsky, I. N. A geometry-based generic predictor for catalytic and allosteric sites. *Protein Eng., Des. Sel.* **2011**, *24* (4), 405–409.
[122] Roche, D. B.; Tetchner, S. J.; McGuffin, L. J. FunFOLD: An improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinf.* **2011**, *12*, 160.
[123] Tan, K. P.; Varadarajan, R.; Madhusudhan, M. S. DEPTH: A web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res.* **2011**, *39* (Web Server issue), W242-8.
[124] Livesay, D. R.; Bahadur, K. C. D.; La, D. Predicting protein functional sites with phylogenetic motifs: Past, present and beyond. In *Protein Function Prediction for Omics Era*; Kihara, D., Ed.; Springer Netherlands: Dordrecht, 2011; pp 93–105.
[125] Neudert, G.; Klebe, G. fconv: Format conversion, manipulation and feature computation of molecular data. *Bioinformatics* **2011**, *27* (7), 1021–1022.
[126] Zhu, H.; Pisabarro, M. T. MSPocket: An orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics* **2011**, *27* (3), 351–358.
[127] Craig, I. R.; Pfleger, C.; Gohlke, H.; Essex, J. W.; Spiegel, K. Pocket-space maps to identify novel binding-site conformations in proteins. *J. Chem. Inf. Model.* **2011**, *51* (10), 2666–2679.
[128] Fukunishi, Y.; Nakamura, H. Prediction of ligand-binding sites of proteins by molecular docking calculation for a random ligand library. *Protein Sci.* **2011**, *20* (1), 95–106.
[129] Hetényi, C.; van der Spoel, D. Toward prediction of functional protein pockets using blind docking and pocket search algorithms. *Protein Sci.* **2011**, *20* (5), 880–893.
[130] Mach, P.; Koehl, P. Geometric measures of large biomolecules: Surface, volume, and pockets. *J. Comput. Chem.* **2011**, *32* (14), 3023–3038.

[131] Zhang, Z.; Li, Y.; Lin, B.; Schroeder, M.; Huang, B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **2011**, *27* (15), 2083–2088.
[132] Zhang, C.; Lai, L. AutoMatch: Target-binding protein design and enzyme design by automatic pinpointing potential active sites in available protein scaffolds. *Proteins* **2012**, *80* (4), 1078–1094.
[133] Giard, J.; Alface, P. R.; Gala, J.-L.; Macq, B. Fast surface-based travel depth estimation algorithm for macromolecule surface shape description. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2011**, *8* (1), 59–68.
[134] Olechnovic, K.; Margelevicius, M.; Venclovas, C. Voroprot: An interactive tool for the analysis and visualization of complex geometric features of protein structure. *Bioinformatics* **2011**, *27* (5), 723–724.
[135] Raunest, M.; Kandt, C. dxTuber: Detecting protein cavities, tunnels and clefts based on protein and solvent dynamics. *J. Mol. Graphics Modell.* **2011**, *29* (7), 895–905.
[136] Schmidtke, P.; Bidon-Chanal, A.; Luque, F. J.; Barril, X. MDpocket: Open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* **2011**, *27* (23), 3276–3285.
[137] Krone, M.; Falk, M.; Rehm, S.; Pleiss, J.; Ertl, T. Interactive exploration of protein cavities. *Comput. Graph. Forum* **2011**, *30* (3), 673–682.
[138] Lindow, N.; Baum, D.; Hege, H.-C. Voronoi-based extraction and visualization of molecular paths. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17* (12), 2025–2034.
[139] Ngan, C.-H.; Hall, D. R.; Zerbe, B.; Grove, L. E.; Kozakov, D.; Vajda, S. FTSite: High accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* **2012**, *28* (2), 286–287.
[140] Schneider, S.; Zacharias, M. Combining geometric pocket detection and desolvation properties to detect putative ligand binding sites on proteins. *J. Struct. Biol.* **2012**, *180* (3), 546–550.
[141] Jenik, M.; Parra, R. G.; Radusky, L. G.; Turjanski, A.; Wolynes, P. G.; Ferreiro, D. U. Protein Frustratometer: A tool to localize energetic frustration in protein molecules. *Nucleic Acids Res.* **2012**, *40* (Web Server issue), W348-51.
[142] Källberg, M.; Wang, H.; Wang, S.; Peng, J.; Wang, Z.; Lu, H.; Xu, J. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **2012**, *7* (8), 1511–1522.
[143] Nemoto, W.; Toh, H. Functional region prediction with a set of appropriate homologous sequences - An index for sequence selection by integrating structure and sequence information with spatial statistics. *BMC Struct. Biol.* **2012**, *12*, 11.
[144] Xie, Z.-R.; Hwang, M.-J. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. *Bioinformatics* **2012**, *28* (12), 1579–1585.
[145] Roy, A.; Zhang, Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* **2012**, *20* (6), 987–997.
[146] Bianchi, V.; Gherardini, P. F.; Helmer-Citterich, M.; Ausiello, G. Identification of binding pockets in protein structures using a knowledge-based potential derived from local structural similarities. *BMC Bioinf.* **2012**, *13 Suppl 4*, S17.
[147] Somarowthu, S.; Ondrechen, M. J. POOL server: Machine learning application for functional site prediction in proteins. *Bioinformatics* **2012**, *28* (15), 2078–2079.
[148] Nadzirin, N.; Gardiner, E. J.; Willett, P.; Artymiuk, P. J.; Firdaus-Raih, M. SPRITE and ASSAM: Web servers for side chain 3D-motif searching in protein structures. *Nucleic Acids Res.* **2012**, *40* (Web Server issue), W380-6.
[149] Kinjo, A. R.; Nakamura, H. Composite structural motifs of binding sites for delineating biological functions of proteins. *PLoS ONE* **2012**, *7* (2), e31437.
[150] Tseng, Y. Y.; Li, W.-H. Classification of protein functional surfaces using structural characteristics. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (4), 1170–1175.

[151] Sridharamurthy, R.; Doraiswamy, H.; Patel, S.; Varadarajan, R.; Natarajan, V. Extraction of robust voids and pockets in proteins. In *Eurovis - Short Papers*; Hlawitschka, M., Weinkauf, T., Eds.; The Eurographics Association, 2013; pp 67–71.
[152] Cammisa, M.; Correra, A.; Andreotti, G.; Cubellis, M. V. Identification and analysis of conserved pockets on protein surfaces. *BMC Bioinf.* **2013**, *14 Suppl 7*, S9.
[153] Brylinski, M.; Feinstein, W. P. eFindSite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J. Comput.-Aided Mol. Des.* **2013**, *27* (6), 551–567.
[154] Grove, L. E.; Hall, D. R.; Beglov, D.; Vajda, S.; Kozakov, D. FTFlex: Accounting for binding site flexibility to improve fragment-based identification of druggable hot spots. *Bioinformatics* **2013**, *29* (9), 1218–1219.
[155] Chen, S.-W. W.; Pellequer, J.-L. Adepth: New representation and its implications for atomic depths of macromolecules. *Nucleic Acids Res.* **2013**, *41* (Web Server issue), W412-6.
[156] Kaya, C.; Armutlulu, A.; Ekesan, S.; Haliloglu, T. MCPath: Monte Carlo path generation approach to predict likely allosteric pathways and functional residues. *Nucleic Acids Res.* **2013**, *41* (Web Server issue), W249-55.
[157] Roche, D. B.; Buenavista, M. T.; McGuffin, L. J. The FunFOLD2 server for the prediction of protein-ligand interactions. *Nucleic Acids Res.* **2013**, *41* (Web Server issue), W303-7.
[158] Yu, D.-J.; Hu, J.; Yang, J.; Shen, H.-B.; Tang, J.; Yang, J.-Y. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2013**, *10* (4), 994–1008.
[159] Wong, G. Y.; Leung, F. H. F.; Ling, S. H. Predicting protein-ligand binding site using support vector machine with protein properties. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2013**, *10* (6), 1517–1529.
[160] Gao, M.; Skolnick, J. APoc: Large-scale identification of similar protein pockets. *Bioinformatics* **2013**, *29* (5), 597–604.
[161] Yang, J.; Roy, A.; Zhang, Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29* (20), 2588–2595.
[162] Lo, Y.-T.; Wang, H.-W.; Pai, T.-W.; Tzou, W.-S.; Hsu, H.-H.; Chang, H.-T. Protein-ligand binding region prediction (PLB-SAVE) based on geometric features and CUDA acceleration. *BMC Bioinf.* **2013**, *14 Suppl 4*, S4.
[163] Yuan, Y.; Pei, J.; Lai, L. Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.* **2013**, *19* (12), 2326–2333.
[164] Lee, H. S.; Im, W. Ligand binding site detection by local structure alignment and its performance complementarity. *J. Chem. Inf. Model.* **2013**, *53* (9), 2462–2470.
[165] Ribeiro, J. V.; Tamames, J. A. C.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J. Volarea - A bioinformatics tool to calculate the surface area and the volume of molecular systems. *Chem. Biol. Drug Des.* **2013**, *82* (6), 743–755.
[166] Sehnal, D.; Svobodová Vařeková, R.; Berka, K.; Pravda, L.; Navrátilová, V.; Banáš, P.; Ionescu, C.-M.; Otyepka, M.; Koča, J. MOLE 2.0: Advanced approach for analysis of biomacromolecular channels. *J. Cheminf.* **2013**, *5* (1), 39.
[167] Parulek, J.; Turkay, C.; Reuter, N.; Viola, I. Visual cavity analysis in molecular simulations. *BMC Bioinf.* **2013**, *14 Suppl 19*, S4.
[168] Krone, M.; Reina, G.; Schulz, C.; Kulschewski, T.; Pleiss, J.; Ertl, T. Interactive extraction and tracking of biomolecular surface features. *Comput. Graph. Forum* **2013**, *32* (3pt3), 331–340.
[169] Lindow, N.; Baum, D.; Bondar, A.-N.; Hege, H.-C. Exploring cavity dynamics in biomolecular systems. *BMC Bioinf.* **2013**, *14 Suppl 19*, S5.

[170] Decherchi, S.; Rocchia, W. A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. *PLoS ONE* **2013**, *8* (4), e59744.
[171] Kokh, D. B.; Richter, S.; Henrich, S.; Czodrowski, P.; Rippmann, F.; Wade, R. C. TRAPP: A tool for analysis of transient binding pockets in proteins. *J. Chem. Inf. Model.* **2013**, *53* (5), 1235–1252.
[172] Wang, K.; Gao, J.; Shen, S.; Tuszynski, J. A.; Ruan, J.; Hu, G. An accurate method for prediction of protein-ligand binding site on protein surface using SVM and statistical depth function. *Biomed Res. Int.* **2013**, *2013*, 409658.
[173] Suplatov, D.; Kirilin, E.; Arbatsky, M.; Takhaveev, V.; Svedas, V. pocketZebra: A web-server for automated selection and classification of subfamily-specific binding sites by bioinformatic analysis of diverse protein families. *Nucleic Acids Res.* **2014**, *42* (Web Server issue), W344-9.
[174] Heo, L.; Shin, W.-H.; Lee, M. S.; Seok, C. GalaxySite: Ligand-binding-site prediction by using molecular docking. *Nucleic Acids Res.* **2014**, *42* (Web Server issue), W210-4.
[175] Saberi Fathi, S. M.; Tuszynski, J. A. A simple method for finding a protein's ligand-binding pockets. *BMC Struct. Biol.* **2014**, *14*, 18.
[176] Benkaidali, L.; André, F.; Maouche, B.; Siregar, P.; Benyettou, M.; Maurel, F.; Petitjean, M. Computing cavities, channels, pores and pockets in proteins from non-spherical ligands models. *Bioinformatics* **2014**, *30* (6), 792–800.
[177] Kim, J.-K.; Cho, Y.; Laskowski, R. A.; Ryu, S. E.; Sugihara, K.; Kim, D.-S. BetaVoid: Molecular voids via β-complexes and Voronoi diagrams. *Proteins* **2014**, *82* (9), 1829–1849.
[178] Oliveira, S. H. P.; Ferraz, F. A. N.; Honorato, R. V.; Xavier-Neto, J.; Sobreira, T. J. P.; de Oliveira, Paulo S L. KVFinder: Steered identification of protein cavities as a PyMOL plugin. *BMC Bioinf.* **2014**, *15*, 197.
[179] Kozlíková, B.; Sebestova, E.; Sustr, V.; Brezovsky, J.; Strnad, O.; Daniel, L.; Bednar, D.; Pavelka, A.; Manak, M.; Bezdeka, M.; Benes, P.; Kotry, M.; Gora, A.; Damborský, J.; Sochor, J. CAVER Analyst 1.0: Graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. *Bioinformatics* **2014**, *30* (18), 2684–2685.
[180] Lindow, N.; Baum, D.; Hege, H.-C. Ligand Excluded Surface: A new type of molecular surface. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20* (12), 2486–2495.
[181] Paramo, T.; East, A.; Garzón, D.; Ulmschneider, M. B.; Bond, P. J. Efficient characterization of protein cavities within molecular simulation trajectories: Trj_cavity. *J. Chem. Theory Comput.* **2014**, *10* (5), 2151–2164.
[182] Krone, M.; Kauker, D.; Reina, G.; Ertl, T. Visual analysis of dynamic protein cavities and binding sites. In *2014 IEEE Pacific Visualization Symposium*; IEEE, 2014 - 2014; pp 301–305.
[183] Mirceva, G.; Kulakov, A. Improvement of protein binding sites prediction by selecting amino acid residues' features. *J. Struct. Biol.* **2015**, *189* (1), 9–19.
[184] Kim, J.-K.; Cho, Y.; Lee, M.; Laskowski, R. A.; Ryu, S. E.; Sugihara, K.; Kim, D.-S. BetaCavityWeb: A webserver for molecular voids and channels. *Nucleic Acids Res.* **2015**, *43* (W1), W413-8.
[185] Hung, L. V.; Caprari, S.; Bizai, M.; Toti, D.; Polticelli, F. LIBRA: LIgand Binding site Recognition Application. *Bioinformatics* **2015**, *31* (24), 4020–4022.
[186] Yu, D.-J.; Hu, J.; Li, Q.-M.; Tang, Z.-M.; Yang, J.-Y.; Shen, H.-B. Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction. *IEEE Trans. Nanobiosci.* **2015**, *14* (1), 45–58.
[187] Gaudreault, F.; Morency, L.-P.; Najmanovich, R. J. NRGsuite: a PyMOL plugin to perform docking simulations in real time using FlexAID. *Bioinformatics* **2015**, *31* (23), 3856–3858.
[188] Czirják, G. PrinCCes: Continuity-based geometric decomposition and systematic visualization of the void repertoire of proteins. *J. Mol. Graphics Modell.* **2015**, *62*, 118–127.

[189] Krivák, R.; Hoksza, D. P2RANK: Knowledge-based ligand binding site prediction using aggregated local features. In *Algorithms for Computational Biology*; Dediu, A.-H., Hernández-Quiroz, F., Martín-Vide, C., Rosenblueth, D. A., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2015; pp 41–52.
[190] Guo, Z.; Li, B.; Cheng, L.-T.; Zhou, S.; McCammon, J. A.; Che, J. Identification of protein-ligand binding sites by the level-set variational implicit-solvent approach. *J. Chem. Theory Comput.* **2015**, *11* (2), 753–765.
[191] Buša, J.; Hayryan, S.; Hu, C.-K.; Wu, M.-C. CAVE-CL: An OpenCL version of the package for detection and quantitative analysis of internal cavities in a system of overlapping balls: Application to proteins. *Comput. Phys. Commun.* **2015**, *190*, 224–227.
[192] Desdouits, N.; Nilges, M.; Blondel, A. Principal component analysis reveals correlation of cavities evolution and functional motions in proteins. *J. Mol. Graphics Modell.* **2015**, *55*, 13–24.
[193] Harigua-Souiai, E.; Cortes-Ciriano, I.; Desdouits, N.; Malliavin, T. E.; Guizani, I.; Nilges, M.; Blondel, A.; Bouvier, G. Identification of binding sites and favorable ligand binding moieties by virtual screening and self-organizing map analysis. *BMC Bioinf.* **2015**, *16*, 93.
[194] Chen, P.; Hu, S.; Zhang, J.; Gao, X.; Li, J.; Xia, J.; Wang, B. A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2016**, *13* (5), 901–912.
[195] Gao, J.; Zhang, Q.; Liu, M.; Zhu, L.; Wu, D.; Cao, Z.; Zhu, R. bSiteFinder, an improved protein-binding sites prediction server based on structural alignment: More accurate and less time-consuming. *J. Cheminf.* **2016**, *8*, 38.
[196] Parra, R. G.; Schafer, N. P.; Radusky, L. G.; Tsai, M.-Y.; Guzovsky, A. B.; Wolynes, P. G.; Ferreiro, D. U. Protein Frustratometer 2: A tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res.* **2016**, *44* (W1), W356-60.
[197] Komiyama, Y.; Banno, M.; Ueki, K.; Saad, G.; Shimizu, K. Automatic generation of bioinformatics tools for predicting protein-ligand binding sites. *Bioinformatics* **2016**, *32* (6), 901–907.
[198] Ravindranath, P. A.; Sanner, M. F. AutoSite: An automated approach for pseudo-ligands prediction-from ligand-binding sites identification to predicting key ligand atoms. *Bioinformatics* **2016**, *32* (20), 3142–3149.
[199] Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; Fraser, J. S.; Sali, A. CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J. Mol. Biol.* **2016**, *428* (4), 709–719.
[200] Roche, D. B.; McGuffin, L. J. *In silico* identification and characterization of protein-ligand binding sites. *Methods Mol. Biol.* **2016**, *1414*, 1–21.
[201] Jian, J.-W.; Elumalai, P.; Pitti, T.; Wu, C. Y.; Tsai, K.-C.; Chang, J.-Y.; Peng, H.-P.; Yang, A.-S. Predicting ligand binding sites on protein surfaces by 3-dimensional probability density distributions of interacting atoms. *PLoS ONE* **2016**, *11* (8), e0160315.
[202] Tsujikawa, H.; Sato, K.; Wei, C.; Saad, G.; Sumikoshi, K.; Nakamura, S.; Terada, T.; Shimizu, K. Development of a protein-ligand-binding site prediction method based on interaction energy and sequence conservation. *J. Struct. Funct. Genomics* **2016**, *17* (2-3), 39–49.
[203] Glantz-Gashai, Y.; Meirson, T.; Samson, A. O. Normal modes expose active sites in enzymes. *PLoS Comput. Biol.* **2016**, *12* (12), e1005293.
[204] Cao, C.; Xu, S. Improving the performance of the PLB index for ligand-binding site prediction using dihedral angles and the solvent-accessible surface area. *Sci. Rep.* **2016**, *6*, 33232.
[205] Enem, Anthony, Porter, Benjamin, Uwacu, Diane, Thomas, Shawna. *Predicting ligand binding sites with UOBPRM and machine learning*.
[206] Kim, B.; Lee, J. E.; Kim, Y. J.; Kim, K.-J. GPU accelerated finding of channels and tunnels for a protein molecule. *Int. J. Parallel Prog.* **2016**, *44* (1), 87–108.

[207] Jurcik, A.; Parulek, J.; Sochor, J.; Kozlíková, B. Accelerated visualization of transparent molecular surfaces in molecular dynamics. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*; IEEE, 2016 - 2016; pp 112–119.
[208] Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; Fabritiis, G. de. DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017**, *33* (19), 3036–3042.
[209] Stank, A.; Kokh, D. B.; Horn, M.; Sizikova, E.; Neil, R.; Panecka, J.; Richter, S.; Wade, R. C. TRAPP webserver: Predicting protein binding site flexibility and detecting transient binding pockets. *Nucleic Acids Res.* **2017**, *45* (W1), W325-W330.
[210] Lee, H. S.; Im, W. G-LoSA for prediction of protein-ligand binding sites and structures. *Methods Mol. Biol.* **2017**, *1611*, 97–108.
[211] Zhang, C.; Freddolino, P. L.; Zhang, Y. COFACTOR: Improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res.* **2017**, *45* (W1), W291-W299.
[212] Singh, K.; Lahiri, T. A new search subspace to compensate failure of cavity-based localization of ligand-binding sites. *Comput. Biol. Chem.* **2017**, *68*, 6–11.
[213] Dias, S. E. D.; Martins, A. M.; Nguyen, Q. T.; Gomes, A. J. P. GPU-based detection of protein cavities using Gaussian surfaces. *BMC Bioinf.* **2017**, *18* (1), 493.
[214] Dias, S. E. D.; Nguyen, Q. T.; Jorge, J. A.; Gomes, A. J. P. Multi-GPU-based detection of protein cavities using critical points. *Future Gener. Comput. Syst.* **2017**, *67*, 430–440.
[215] Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47* (12), 2977–2980.
[216] Zhang, Y.; Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33* (7), 2302–2309.
[217] Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.* **2012**, *52* (2), 360–372.
[218] Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins* **2005**, *60* (3), 333–340.
[219] Chen, X.; Liu, M.; Gilson, M. K. BindingDB: A web-accessible molecular recognition database. *Comb. Chem. High Throughput Screen.* **2001**, *4* (8), 719–725.
[220] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.
[221] Huang, W.; Wang, G.; Shen, Q.; Liu, X.; Lu, S.; Geng, L.; Huang, Z.; Zhang, J. ASBench: Benchmarking sets for allosteric discovery. *Bioinformatics* **2015**, *31* (15), 2598–2600.

APPENDIX 2: Binding Site Comparison

Table A2.1. Dataset of structures with identical sequences (datasets 1/1.2). Structures of dataset 1.2 are given in bold characters. For proteins from an organism other than *Homo sapiens*, the name is given in brackets.

PDB-ID	Chain	Ligand-ID	PDB-ID	Chain	Ligand-ID	PDB-ID	Chain	Ligand-ID
thrombin			carbonic anhydrase II			heat shock protein 90		
2zda	**H**	**32U**	1bnq	A	AL4	**1uye**	**A**	**PU9**
2zc9	**H**	**22U**	1bnw	A	TPD	**1uyd**	**A**	**PU8**
3qwc	H	98P	1bnv	A	AL7	1uyg	A	PU2
3sv2	**H**	**P05**	1bnu	A	AL3	1uyf	A	PU1
3si3	**H**	**B03**	1bnn	A	AL1	1uyc	A	PU7
3si4	**H**	**B04**	1bnm	A	AL8	1uyh	A	PU0
1ypl	H	RA8	3oys	A	OYS	2xhr	A	C0P
1ypm	H	RA4	1i91	A	INQ	2qfo	A	A51
1ype	H	UIP	1okl	A	MNS	2xk2	A	ADP
2uuj	B	896	1bn1	A	AL5	2xht	A	C0Y
2zff	**H**	**53U**	4bf6	A	X0Q	2h55	A	DZ8
3rmm	H	M32	3k2f	A	NKX	4nh8	A	2LC
2znk	H	31U	3m98	A	E02	2uwd	A	2GG
2zfp	H	19U	**3m96**	**A**	**E38**	3wha	B	WHA
2zfq	H	45U	3m04	A	BE9	3vha	A	VHA
2fes	H	3SP	3bet	A	CTF	**1uy7**	**A**	**PU4**
2feq	H	34P	3n0n	A	P9B	**1uy6**	**A**	**PU3**
3p17	**H**	**99P**	**3s9t**	**A**	**E49**	1uyk	A	PUX
3qx5	H	02P	4dz9	A	ID4	4efu	A	EFU
3t5f	H	M34	2h15	A	B19	1uy9	A	PU6
2anm	H	CDO	3d9z	A	D9Z	3r4p	A	FU7
1d6w	A	00R	2hl4	A	BO1	3r4o	B	FU3
3u98	H	BJA	4fpt	A	0VZ	3r4n	B	FU5
2ank	H	N12	3p5l	A	IT5	1osf	A	KOS
2zf0	**H**	**51U**	2aw1	A	COX	3inx	A	JZC
2a2x	H	NA9	1cim	A	PTS	3inw	A	JZB
1kts	B	C24	2eu2	A	5DS	3hz5	A	Z64
1ktt	B	C02	4mty	A	SBW	2byh	A	2D7
1o0d	H	163	1xq0	A	4TR	3mnr	P	SD1
3da9	B	44U	1a42	A	BZU	2xab	A	VHD
1nzq	H	162	2wd3	A	MS4	3bmy	A	CXZ
2cf9	H	348	3mhl	A	J71	2vci	A	2GJ
4e7r	G	0NW	3mhm	A	J75	2vcj	A	2EQ
3rm0	H	S54	3mhc	A	ARZ	2yi7	A	BZ8
3rm2	H	S00	2hd6	A	BOS	3rlr	A	3RR
3rlw	H	S28	3dd8	A	2C7	3b28	A	B2X
3rly	H	S29	3mzc	A	S6I	3b25	A	B2K
2jh0	D	701	3ffp	X	LC1	2fwy	A	H64
2jh6	D	894	3f4x	A	KLT	2fwz	A	H71
3rml	H	M31	3l14	A	I7B	4b7p	A	9UN
3u9a	H	S33	3t84	A	SG6	3o0i	A	P54
3rmn	H	M41	4kni	A	E1E	2qg0	A	A94
3rmo	H	S04	4knj	A	E1F	4bqg	A	50Q
1d9i	A	00P	3vbd	A	0FZ	4egk	A	RDC
2v3o	H	I26	1ze8	A	PIU	3t1k	A	ANP
1c4u	2	IH1	1kwq	A	SG1	4bqj	A	XKL
1c4v	2	IH2	3r17	B	5UM	4jql	A	VJ6
1ypk	H	CCR	3m2x	A	BEX	2wi6	A	ZZ6
3shc	H	B01	1z9y	A	FUN	2wi4	A	ZZ4

Table A2.1 ***(continued)*****. Dataset of structures with identical sequences (datasets 1/1.2).** Structures of dataset 1.2 are given in bold characters. For proteins from an organism other than *Homo sapiens*, the name is given in brackets.

PDB-ID	Chain	Ligand-ID	PDB-ID	Chain	Ligand-ID	PDB-ID	Chain	Ligand-ID
thrombin			carbonic anhydrase II			heat shock protein 90		
3sha	H	P97	1i8z	A	INL	3qdd	A	94M
1d4p	B	BPP	4e3f	A	GRE	2byi	A	2DD
2ziq	H	26U	3sax	A	E50	1uy8	A	PU5
3biu	H	10U	1ttm	A	667	3t10	A	ACP
2uuk	B	897	3c7p	A	POF	3k98	B	1RC
2gde	H	SN3	2hoc	A	1CN	3d0b	A	SNX
3qto	**H**	**10P**	**3m67**	**A**	**E36**	3k97	A	4CD
3qtv	**H**	**06P**	3sbi	A	E90			
1oyt	H	FSN	3igp	A	DT7			
1d3d	B	BZT	4itp	A	1GD			
2zi2	H	24U	3po6	A	DT9			
1d3p	B	BT3	3b4f	A	TUO			
1g30	B	T87						
1g32	B	R11						
dihydrofolate reductase			isopenicillin N synthetase			macrophage metalloelastase		
3nu0	A	3TU	2bu9	A	HFV	**1rmz**	**A**	**NGH**
3ghw	A	GHW	**2y6f**	**A**	**M9F**	3lk8	A	Z79
3ntz	A	3TZ	**1qiq**	**A**	**ACC**	**3nx7**	**A**	**NHK**
1mvs	A	DTM	**1odm**	**A**	**ASV**	3ts4	A	EEG
1pd9	A	CO4	1odn	A	APV	3rtt	A	KLH
4keb	A	1QZ	**1bk0**	**A**	**ACV**	4guy	A	KLJ
1yho	A	TRR	1w3x	A	W2X	4h76	A	10B
3nxt	A	D2E	**1w05**	**A**	**W05**	3n2v	A	JT5
4kd7	A	9DR	1w3v	A	MDZ	4gr0	A	R4B
1hfr	A	MOT	**2vbp**	**A**	**VB1**	4efs	A	E37
1u72	A	MTX	3zku	A	HCV	**3f15**	**A**	**HS1**
4m6l	**A**	**21V**	3zky	A	WT4	3f16	A	HS3
1s3w	A	TQT	3zoi	A	M2W	3f17	A	HS4
1s3v	A	TQD	**2y60**	**A**	**M8F**	3f18	A	HS5
4ddr	A	MMV	1qjf	A	ACS	2hu6	A	37A
2dhf	**A**	**DZF**	1hb4	A	SCV	3ehx	A	BDL
3nxv	A	D2F	2ivi	B	ACW	3ehy	A	TBL
3nxy	A	D2H	2ivj	A	BCV	3rts	A	KLG
1drf	A	FOL	1hb1	A	OCV	3tsk	A	QEG
4g95	A	OAG	2vcm	A	M11			
1ohk	A	COP	2vbd	A	V10			
1kmv	A	LII	**1uzw**	**A**	**CDH**			
3s7a	A	684	2vau	A	V20			
4kbn	A	25U	4bb3	A	KKA			
leukotriene A4 hydrolase			D-alanyl-D-alanine carboxypeptidase (*Streptomyces* sp. R61)			Bromodomain-containing protein 4		
3u9w	A	28P	1pwl	A	BFI	**4j0r**	**A**	**1H2**
3b7u	X	KEL	3g5e	A	Q74	**4j0s**	**A**	**1H3**
3b7r	L	BIR	2qxw	A	LDT	4men	A	25K
3chs	A	4BU	2ikh	A	LIT	4meo	A	25V
3chp	**A**	**4BO**	**1x97**	**A**	**FIR**	4hxp	A	1A5
2r59	A	PH0	3p2v	A	DOY	4hxr	A	1A4
3cho	**A**	**4BG**	3t42	A	3T4	4a9l	A	P9L
3fh8	A	27P	2nvd	A	ITB	4gpj	A	0Q1

Table A2.1 ***(continued)*****. Dataset of structures with identical sequences (datasets 1/1.2).** Structures of dataset 1.2 are given in bold characters. For proteins from an organism other than *Homo sapiens*, the name is given in brackets.

PDB-ID	**Chain**	**Ligand-ID**	**PDB-ID**	**Chain**	**Ligand-ID**	**PDB-ID**	**Chain**	**Ligand-ID**
leukotriene A4 hydrolase			D-alanyl-D-alanine carboxypeptidase (*Streptomyces* sp. R61)			Bromodomain-containing protein 4		
3fh7	A	25P	2nvc	A	ITA	4c67	A	L5S
3fh5	A	24P	3u2c	A	SUZ	3u5l	A	08K
3chq	**A**	**4BQ**	3dn5	A	53N	3u5j	A	08H
3ful	A	52D	2ikg	A	BTO	4hbx	A	14X
3fum	A	80A	1z3n	A	3NA	4hbv	A	15E
3fun	A	798	**1pwm**	**A**	**FID**			
3fui	A	812	1z89	A	62P			
3fuk	A	58Z	1t41	A	ID5			
3fuf	A	BES	4gca	A	2X9			
tankyrase-2			angiotensin-converting enzyme (*Drosophila melanogaster*)			pantothenate kinase (*Mycobacterium tuberculosis*)		
3u9y	A	09L	4ca7	A	3EF	**2zsa**	**A**	**ADP**
3p0n	C	BPU	2x95	A	X95	2zsd	A	COA
4bud	A	29F	**2x94**	**A**	**X94**	4bft	B	ZVT
3kr8	C	XAV	2x97	A	RX4	4bfv	B	ZVV
4bux	A	F35	2x96	A	RX3	4bfw	B	ZVW
4buu	**A**	**F38**	2x91	A	LPR	4bfx	A	ZVX
4but	**A**	**31F**	2x90	A	EAL	4bfy	B	ZVY
4buw	A	F33	**2x93**	**A**	**X93**	4bfz	A	ZVZ
4bus	A	32F	**2x92**	**A**	**X92**	**3af2**	**A**	**ACP**
3ua9	B	IWR	2xhm	A	K26	**3af3**	**A**	**GCP**
4j3m	B	AJ8						
4buy	A	F37						

Table A2.2. Dataset of NMR structure ensembles (dataset 2).

PDB-ID	**Chain**	**Ligand-ID**	**|conformers|**	**Protein**
1cz2	A	E2P	12	nonspecific lipid transfer protein (*Triticum aestivum*)
1diu	A	BDM	18	dihydrofolate reductase (*Lactobacillus casei*)
1eio	A	GCH	5	ileal lipid binding protein (*Sus scrofa*)
1j5i	A	NCZ	44	neocarzinostatin (*Streptomyces carzinostaticus*)
1kgl	A	RTL	20	cellular retinol-binding protein type I (*Rattus norvegicus*)
1mux	A	WW7	30	calmodulin (*Xenopus laevis*)
1t84	A	WSK	20	Wiskott-Aldrich syndrome protein (*Homo sapiens*)
1tvc	A	FDA	10	methane monooxygenase component C (*Methylococcus capsulatus*)
1yho	A	TRR	25	dihydrofolate reductase (*Homo sapiens*)
2jt2	A	C90	25	UDP-3-O-[3-hydroxymyristoyl]N-acetylglucosamine deacetylase (*Aquifex aeolicus*)
2k31	A	35G	20	cGMP-specific 3',5'-cyclic phosphodiesterase (*Mus musculus*)

Table A2.2 *(continued)*. Dataset of NMR structure ensembles (dataset 2).

PDB-ID	Chain	Ligand-ID	|conformers|	Protein
2k5t	A	COA	20	PanD maturation factor (*Escherichia coli*)
2l0x	A	GDP	20	GTP-binding protein Rheb (*Rattus norvegicus*)
2l2s	A	L2S	20	peptidyl-prolyl cis-trans isomerase (*Burkholderia pseudomallei*)
2l8r	A	APR	20	O-acetyl-ADP-ribose deacetylase 1 (*Homo sapiens*)
2lzg	A	13Q	5	E3 ubiquitin-protein ligase Mdm2 (*Homo sapiens*)
2z2d	A	HSI	15	macrophage metalloelastase (*Homo sapiens*)

Table A2.3. The Kahraman dataset (dataset 6).

PDB-ID	Chain	Ligand-ID	PDB-ID	Chain	Ligand-ID
12as	A	AMP	1d0c	A	HEM
1amu	A	AMP	1d7c	A	HEM
1c0a	A	AMP	1dk0	A	HEM
1ct9	A	AMP	1eqg	A	HEM
1jp4	A	AMP	1ew0	A	HEM
1kht	B	AMP	1gwe	A	HEM
1qb8	A	AMP	1iqc	A	HEM
1tb7	B	AMP	1naz	A	HEM
8gpb	A	AMP	1np4	A	HEM
1a0i	A	ATP	1po5	A	HEM
1a49	A	ATP	1pp9	C	HEM
1ayl	A	ATP	1qhu	A	HEM
1b8a	A	ATP	1qla	C	HEM
1dv2	A	ATP	1qpa	A	HEM
1dy3	A	ATP	1sox	A	HEM
1e2q	A	ATP	2cpo	A	HEM
1e8x	A	ATP	1ej2	A	NAD
1esq	A	ATP	1hex	A	NAD
1gn8	A	ATP	1ib0	A	NAD
1kvk	A	ATP	1jq5	A	NAD
1o9t	B	ATP	1mew	A	NAD
1rdq	E	ATP	1mi3	A	NAD
1tid	A	ATP	1o04	A	NAD
1e3r	A	AND	1og3	A	NAD
1j99	A	AND	1qax	B	NAD
1fds	A	EST	1rlz	A	NAD
1lhu	A	EST	1s7g	B	NAD
1qkt	A	EST	1t2d	A	NAD
1cqx	A	FAD	1tox	A	NAD
1e8g	A	FAD	2a5f	B	NAD
1evi	A	FAD	2npx	A	NAD
1h69	A	FAD	1a6q	A	PO4
1hsk	A	FAD	1b8o	A	PO4
1jqi	A	FAD	1brw	A	PO4
1jr8	B	FAD	1cqj	B	PO4
1k87	A	FAD	1d1q	B	PO4
1pox	A	FAD	1dak	A	PO4
3grs	A	FAD	1e9g	A	PO4

Table A2.3 *(continued)*. The Kahraman dataset (dataset 6).

PDB-ID	Chain	Ligand-ID	PDB-ID	Chain	Ligand-ID
1dnl	A	FMN	1ejd	B	PO4
1f5v	A	FMN	1euc	A	PO4
1ja1	A	FMN	1ew2	A	PO4
1mvl	A	FMN	1fbt	B	PO4
1p4c	A	FMN	1gyp	A	PO4
1p4m	A	FMN	1h6l	A	PO4
1bdg	A	GLC	1ho5	B	PO4
1cq1	A	GLC	1l5w	B	PO4
1k1w	A	GLC	1l7m	A	PO4
1nf5	B	GLC	1lby	A	PO4
2gbp	A	GLC	1lyv	A	PO4
			1qf5	A	PO4
			1tco	A	PO4

Table A2.4. The Barelier dataset (dataset 7). The RMSD values were obtained by a least-squares fitting of corresponding ligand atoms of both protein-ligand complexes with UCSF Chimera[1].

PDB-ID 1	Ligand 1 (Name. Chain. ID)	Cofactors 1	PDB-ID 2	Ligand 2 (Name. Chain. ID)	Cofactors 2	Pair Class	Ligand RMSD [Å]
2rh1	CAU.A.408	-	2q6h	CXX.A.801	-	A	3.755
2b50	VCA.A.1001	-	4bvm	VCA.A.1133	-	A	2.815
1vyg	ACD.A.1134	-	1diy	ACD.A.700	-	A	2.808
1mv9	HXA.A.200	-	1fdq	HXA.A.133	-	A	2.517
2ans	2AN.A.201	-	1eyn	2AN.A.550	-	A	1.526
2wd9	IBP.A.1570	-	3p6h	IBP.A.133	-	A	1.442
4nyq	EIC.A.208	-	4ia6	EIC.A.608	-	A	1.424
3hig	BRN.A.901	-	2gby	BRN.A.1110	-	A	1.360
2qre	AMZ.G.1001	-	1m9n	AMZ.A.1001	-	A	1.302
1eyn	2AN.A.550	-	1ow4	2AN.B.1003	-	A	1.043
3cf9	AGI.A.161	-	4hkk	AGI.A.1204	-	A	1.010
3fty	3IP.A.710	-	1w7h	3IP.A.1355	-	A	0.909
2ans	2AN.A.201	-	1ow4	2AN.B.1003	-	A	0.866
2x8z	X8Z.A.1615	-	4dpr	X8Z.A.702	-	A	0.854
2wyd	DAO.B.1552	-	2ovd	DAO.A.190	-	A	0.849
3bra	AEF.A.394	-	4n7c	AEF.A.202	-	A	0.822
2g0l	FLN.A.114	-	4hki	FLN.A.1204	-	A	0.784
1e6q	NTZ.M.999	-	1nok	NTZ.A.998	-	A	0.266
2ph9	GNT.A.301	-	1dx6	GNT.A.1537	PG4	A	0.234
1tx0	PT1.A.283	SO4	1br6	PT1.A.301	-	B	2.107
1rf7	DHF.A.161	-	1lcb	DHF.A.318	TMP	B	1.861
3hbf	MYC.A.901	-	2o63	MYC.A.501	IMD	B	1.756
2hkk	ALE.A.300	-	3pah	XDE.A.700	-	B	1.730
1z9y	FUN.A.500	-	2xn5	FUN.A.1356	-	B	1.676
4hcy	0YN.A.401	-	4mhw	0YN.A.201	-	B	1.641
1ve7	4NP.B.1701	-	2i6p	4NP.A.401	-	B	1.503
2v6t	H2B.A.1104	-	2fbz	H2B.X.903	HEM	B	1.491
3hlg	LVA.A.500	-	1t02	LVA.A.429	-	B	1.366
1pa9	CSN.A.737	-	1e2s	CSN.P.701	-	B	1.231
2nsl	C2R.A.300	-	2gqs	C2R.A.242	FMT	B	1.201
3ln1	CEL.A.682	-	1oq5	CEL.A.701	-	B	1.182
3wl8	OCA.A.402	-	2wg9	OCA.A.301	-	B	1.143
2vyv	1GP.D.1334	NAD	3ew5	1GP.B.11	-	B	1.048
2yyj	4HP.A.550	FAD	2yaj	4HP.A.1898	-	B	0.878

Table A2.4 *(continued)*. The Barelier dataset (dataset 7). The RMSD values were obtained by a least-squares fitting of corresponding ligand atoms of both protein-ligand complexes with UCSF Chimera[1].

PDB-ID 1	Ligand 1 (Name. Chain. ID)	Cofactors 1	PDB-ID 2	Ligand 2 (Name. Chain. ID)	Cofactors 2	Pair Class	Ligand RMSD [Å]
4kcx	1QK.A.201	-	4kd1	1QK.A.302	-	B	0.870
3hbg	MTE.A.501	-	3etr	MTE.C.1326	MOS	B	0.796
3ebp	CPB.A.940	-	3blr	CPB.A.940	-	B	0.765
3eau	PDN.A.501	NDP	2q1v	PDN.A.248	-	B	0.751
4lzr	LOC.A.201	-	4o2b	LOC.B.503	-	B	0.673
1mxh	DHF.A.1278	NAP	1rf7	DHF.A.161	-	B	0.667
1xql	4AX.B.505	-	1pb9	4AX.A.901	-	B	0.593
1b66	BIO.A.700	-	1sep	BIO.A.801	NAP	B	0.543
2oyf	IAC.A.1001	-	4hm0	IAC.A.505	-	B	0.523
1eyq	NAR.A.501	-	2uxu	NAR.A.1211	-	B	0.496
1mxh	DHF.A.1278	NAP	1lcb	DHF.A.318	TMP	B	0.188
3dds	CFF.B.903	-	2a3b	CFF.A.1433	-	B	0.067
3ed0	EMO.C.161	-	3bqc	EMO.A.400	-	B	0.055
4cus	ES1.A.2972	-	3kpu	ES1.A.290	-	B	0.035
3b00	16A.A.300	-	3p73	16A.A.274	-	C	2.964
1qhy	CLM.A.999	-	4cla	CLM.A.221		C	2.059
3tvl	3PO.A.231	-	4odj	3PO.A.501	-	C	1.738
1s2c	FLF.A.2001	NAP	2pix	FLF.A.2386	-	C	1.349
2uy4	AZM.A.1311	-	3hs4	AZM.A.701	GOL, ZN	C	1.003
2otf	2TN.A.201	-	2nuv	2TN.A.2001	-	C	0.959
2o73	2AL.B.1001	-	2fxl	2AL.A.302	-	C	0.955
3t4k	EMU.A.400	-	4jhi	EMU.A.202		C	0.923
3gcl	AIN.A.609	-	1oxr	AIN.A.141	-	C	0.613
2p1q	IAC.B.901	-	4hm0	IAC.A.505	-	C	0.539
2qvd	BER.A.1811	-	3d6y	BER.A.964		C	0.536
2p1q	IAC.B.901	-	2oyf	IAC.A.1001	-	C	0.433
2fu7	PHN.A.410	-	2lig	PHN.A.390	-	C	0.127
1s4m	LUM.A.294	-	1he5	LUM.A.1207	NAP	C	0.080

Table A2.5. The dataset of successful applications (dataset 8). The reference to the orignal publication of the binding site similarities is given for all structure groups.

PDB-ID	Chain	Ligand (Name.ID.Chain)	Reference
2ejr	A	F2N.1.A	[2]
2bxr	A	FAD.600.A	
1gos	A	FAD.600.A	
2f6d	A	ACR.995.A	[3]
1k1y	A	ACR.660.A	
3lj3	A	WYE.1109.A	[3]
3ma3	A	01I.313.A	
1rx7	A	FOL.161.A	[3]
3bhr	A	THG.266.A	
1dg5	A	NDP.200.A	[3]
3dl6	A	NDP.606.A	
6cox	A	S58.701.A	[4]
1bn4	A	AL9.555.A	
4tzk	A	641.501.A	[5]
2cl5	A	BIE.1218.A	
1eve	A	E20.2001.A	[6]
3cjg	A	KIM.1167.A	

Table A2.5 ***(continued)*****. The dataset of successful applications (dataset 8).** The reference to the orignal publication of the binding site similarities is given for all structure groups.

PDB-ID	Chain	Ligand (Name.ID.Chain)	Reference
2oht	A	IP6.601.A	[7]
2ito	A	IRE.2020.A	
1ohr	A	1UN.201.A	[8]
2j6m	A	AEE.2021.A	
3c0z	B	SHH.301.B	[9]
5hfa	A	FP1.607.A	
3a99	A	AGS.800.A	[10]
1aux	A	ANP.501.A	
1m17	A	AQ4.999.A	[11]
1y57	A	MPZ.600.A	
1qpe	A	PP2.1904.A	
2f4j	A	VX6.514.A	
2rew	A	REW.469.A	[12]
3rr3	A	FLR.700.A	
1xpc	A	AIT.600.A	[13]
2agv	A	TG1.1003.A	
1i7d	A	DG.702.B	[14]
5dlj	D	DT.35.H	
2mjp	A	ANP.500.A	[15]
1a49	A	ATP.535.A	
4kam	A	SO4.501.A	[16]
1e5f	A	PLP.405.A	
1xel	A	NAD.340.A	[17]
1e6w	A	NAD.301.A	
1gal	A	FAD.600.A	
1uk4	B	chain H (pentapeptide)	[17]
1cqq	A	AG7.501.A	
1b0u	A	ATP.301.A	[18]
1jb1	A	PO4.400.A	
1gvr	A	FMN.401.A	[19]
1z41	A	FMN.1500.A	
2hs6	A	FMN.401.A	

Table A2.6. AUC and EFs of different binding site comparison methods for dataset 1.

Method	AUC	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	0.98	7.89	7.91	7.91	7.91	7.91	7.91	7.91
FuzCav	0.94	7.89	7.91	7.91	7.91	7.91	7.91	7.91
FuzCav (PDB)	0.94	7.89	7.91	7.91	7.91	7.91	7.91	7.91
Grim	0.69	7.89	7.91	7.76	7.78	7.38	6.53	5.74
Grim (PDB)	0.62	6.25	6.76	5.51	4.47	4.04	3.59	3.31
IsoMIF	0.77	7.89	7.91	7.83	7.49	7.08	6.61	6.14
KRIPO	0.91	7.89	7.91	7.91	7.91	7.91	7.90	7.89
PocketMatch	0.82	7.59	7.85	7.88	7.90	7.90	7.90	7.88
ProBiS	1.00	7.89	7.91	7.91	7.91	7.91	7.91	7.91
RAPMAD	0.85	7.89	7.91	7.91	7.91	7.79	7.45	6.87
Shaper	0.96	7.89	7.91	7.91	7.91	7.91	7.91	7.91
Shaper (PDB)	0.96	7.89	7.91	7.91	7.91	7.91	7.91	7.91
VolSite/ Shaper	0.93	7.89	7.91	7.91	7.91	7.91	7.91	7.91
VolSite/ Shaper (PDB)	0.94	7.89	7.91	7.91	7.91	7.91	7.91	7.91
SiteAlign	0.97	7.89	7.91	7.91	7.91	7.91	7.91	7.91
SiteEngine	0.96	7.89	7.91	7.91	7.91	7.91	7.91	7.91
SiteHopper	0.98	7.89	7.91	7.91	7.91	7.91	7.91	7.91
SMAP	1.00	7.89	7.91	7.91	7.91	7.91	7.91	7.91
TIFP	0.66	7.89	7.22	5.37	4.37	4.00	3.71	3.40
TIFP (PDB)	0.55	1.64	5.51	4.10	2.88	2.37	2.16	2.05
TM-align	1.00	7.89	7.91	7.91	7.91	7.91	7.91	7.91

Table A2.7. AUC and EFs of different binding site comparison methods for dataset 1.2.

Method	AUC	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	0.91	8.30	8.30	8.30	8.30	8.30	8.40	8.38
FuzCav	0.99	8.30	8.30	8.30	8.30	8.30	8.40	8.38
FuzCav (PDB)	0.99	8.30	8.30	8.30	8.30	8.30	8.40	8.38
Grim	0.97	8.30	8.30	8.30	8.30	8.30	8.40	8.38
Grim (PDB)	0.83	8.30	8.30	6.64	6.22	6.92	7.37	7.30
IsoMIF	0.97	8.30	8.30	8.30	8.30	8.30	8.40	8.38
KRIPO	1.00	8.30	8.30	8.30	8.30	8.30	8.40	8.38
PocketMatch	0.98	8.30	8.30	8.30	8.30	8.30	8.40	8.38
ProBiS	1.00	8.30	8.30	8.30	8.30	8.30	8.40	8.38
RAPMAD	0.83	8.30	8.30	8.30	8.30	8.30	8.40	7.88
Shaper	0.93	8.30	8.30	8.30	8.30	8.30	8.40	8.38
Shaper (PDB)	0.93	8.30	8.30	8.30	8.30	8.30	8.40	8.38
VolSite/ Shaper	0.99	8.30	8.30	8.30	8.30	8.30	8.40	8.38
VolSite/ Shaper (PDB)	1.00	8.30	8.30	8.30	8.30	8.30	8.40	8.38
SiteAlign	1.00	8.30	8.30	8.30	8.30	8.30	8.40	8.38
SiteEngine	1.00	8.30	8.30	8.30	8.30	8.30	8.40	8.38
SiteHopper	0.94	8.30	8.30	8.30	8.30	8.30	8.40	8.38
SMAP	1.00	8.30	8.30	8.30	8.30	8.30	8.40	8.38
TIFP	0.90	8.30	8.30	8.30	8.30	8.30	7.37	7.39
TIFP (PDB)	0.74	8.30	8.30	8.30	8.30	7.75	7.37	6.31
TM-align	1.00	8.30	8.30	8.30	8.30	8.30	8.40	8.38

Table A2.8. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 1. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.97 - 0.98	0.94 - 0.95	0.94 - 0.95	0.61 - 0.62	0.69 - 0.70	0.76 - 0.77	0.91 - 0.92	0.82 - 0.83	1.00 - 1.00	0.84 - 0.85
Cavbase	0.00	-0.03	-0.03	-0.36	-0.28	-0.21	-0.06	-0.15	0.02	-0.13
	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav (PDB)	0.03	0.00	0.00	-0.33	-0.25	-0.18	-0.03	-0.12	0.06	-0.10
	0.00	1.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav	0.03	0.00	0.00	-0.33	-0.25	-0.18	-0.03	-0.12	0.06	-0.10
	0.00	0.95	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim (PDB)	0.36	0.33	0.33	0.00	0.07	0.15	0.30	0.21	0.38	0.23
	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim	0.28	0.25	0.25	-0.07	0.00	0.07	0.22	0.13	0.31	0.15
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
IsoMIF	0.21	0.18	0.18	-0.15	-0.07	0.00	0.15	0.06	0.23	0.08
	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
KRIPO	0.06	0.03	0.03	-0.30	-0.22	-0.15	0.00	-0.09	0.09	-0.07
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
PocketMatch	0.15	0.12	0.12	-0.21	-0.13	-0.06	0.09	0.00	0.18	0.02
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
ProBiS	-0.02	-0.06	-0.06	-0.38	-0.31	-0.23	-0.09	-0.18	0.00	-0.15
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
RAPMAD	0.13	0.10	0.10	-0.23	-0.15	-0.08	0.07	-0.02	0.15	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
VolSite/ Shaper (PDB)	0.04	0.01	0.01	-0.32	-0.24	-0.17	-0.02	-0.11	0.06	-0.09
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper	0.04	0.01	0.01	-0.32	-0.24	-0.17	-0.02	-0.11	0.07	-0.09
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper (PDB)	0.02	-0.02	-0.02	-0.34	-0.27	-0.19	-0.04	-0.14	0.04	-0.11
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper	0.02	-0.02	-0.01	-0.34	-0.27	-0.19	-0.04	-0.14	0.04	-0.11
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteAlign	0.01	-0.02	-0.02	-0.35	-0.28	-0.20	-0.05	-0.14	0.03	-0.12
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteEngine	0.01	-0.02	-0.02	-0.35	-0.27	-0.20	-0.05	-0.14	0.04	-0.12
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteHopper	-0.01	-0.04	-0.04	-0.37	-0.29	-0.22	-0.07	-0.16	0.02	-0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SMAP	-0.02	-0.06	-0.06	-0.38	-0.31	-0.23	-0.09	-0.18	0.00	-0.15
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00
TIFP (PDB)	0.43	0.39	0.39	0.07	0.14	0.22	0.37	0.27	0.45	0.30
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TIFP	0.32	0.28	0.28	-0.04	0.03	0.10	0.25	0.16	0.34	0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TM-align	-0.02	-0.06	-0.06	-0.38	-0.31	-0.23	-0.08	-0.18	0.00	-0.15
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A2.8 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 1. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.93 - 0.94	0.93 - 0.94	0.96 - 0.96	0.96 - 0.96	0.97 - 0.97	0.96 - 0.97	0.98 - 0.99	1.00 - 1.00	0.54 - 0.55	0.66 - 0.67	1.00 - 1.00
Cavbase	-0.04	-0.04	-0.02	-0.02	-0.01	-0.01	0.01	0.02	-0.43	-0.32	0.02
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav (PDB)	-0.01	-0.01	0.02	0.02	0.02	0.02	0.04	0.06	-0.39	-0.28	0.06
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav	-0.01	-0.01	0.02	0.01	0.02	0.02	0.04	0.06	-0.39	-0.28	0.06
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim (PDB)	0.32	0.32	0.34	0.34	0.35	0.35	0.37	0.38	-0.07	0.04	0.38
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim	0.24	0.24	0.27	0.27	0.28	0.27	0.29	0.31	-0.14	-0.03	0.31
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IsoMIF	0.17	0.17	0.19	0.19	0.20	0.20	0.22	0.23	-0.22	-0.10	0.23
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KRIPO	0.02	0.02	0.04	0.04	0.05	0.05	0.07	0.09	-0.37	-0.25	0.08
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pocket-Match	0.11	0.11	0.14	0.14	0.14	0.14	0.16	0.18	-0.27	-0.16	0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ProBiS	-0.06	-0.07	-0.04	-0.04	-0.03	-0.04	-0.02	0.00	-0.45	-0.34	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00
RAPMAD	0.09	0.09	0.11	0.11	0.12	0.12	0.14	0.15	-0.30	-0.18	0.15
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper (PDB)	0.00	0.00	0.02	0.02	0.03	0.03	0.05	0.06	-0.39	-0.27	0.06
	1.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper	0.00	0.00	0.02	0.02	0.03	0.03	0.05	0.07	-0.39	-0.27	0.06
	0.91	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper (PDB)	-0.02	-0.02	0.00	0.00	0.01	0.01	0.02	0.04	-0.41	-0.30	0.04
	0.00	0.00	1.00	0.81	0.00	0.01	0.00	0.00	0.00	0.00	0.00
Shaper	-0.02	-0.02	0.00	0.00	0.01	0.01	0.02	0.04	-0.41	-0.30	0.04
	0.00	0.00	0.81	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteAlign	-0.03	-0.03	-0.01	-0.01	0.00	0.00	0.02	0.03	-0.42	-0.31	0.03
	0.00	0.00	0.00	0.00	1.00	0.05	0.00	0.00	0.00	0.00	0.00
SiteEngine	-0.03	-0.03	-0.01	-0.01	0.00	0.00	0.02	0.04	-0.42	-0.30	0.04
	0.00	0.00	0.01	0.00	0.05	1.00	0.00	0.00	0.00	0.00	0.00
SiteHopper	-0.05	-0.05	-0.02	-0.02	-0.02	-0.02	0.00	0.02	-0.43	-0.32	0.02
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
SMAP	-0.06	-0.07	-0.04	-0.04	-0.03	-0.04	-0.02	0.00	-0.45	-0.34	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NA	0.00	0.00	0.00
TIFP (PDB)	0.39	0.39	0.41	0.41	0.42	0.42	0.43	0.45	0.00	0.11	0.45
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
TIFP	0.27	0.27	0.30	0.30	0.31	0.30	0.32	0.34	-0.11	0.00	0.34
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
TM-align	-0.06	-0.06	-0.04	-0.04	-0.03	-0.04	-0.02	0.00	-0.45	-0.34	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table A2.9. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 1.2. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.87 - 0.94	0.99 - 1.00	0.99 - 1.00	0.79 - 0.87	0.95 - 0.99	0.95 - 0.98	0.99 - 1.00	0.97 - 1.00	1.00 - 1.00	0.80 - 0.87
Cavbase	0.00	0.08	0.08	-0.08	0.06	0.06	0.09	0.07	0.09	-0.08
	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav (PDB)	-0.08	0.00	0.00	-0.17	-0.02	-0.02	0.00	-0.01	0.01	-0.16
	0.00	1.00	1.00	0.00	0.02	0.00	0.43	0.16	0.02	0.00
FuzCav	-0.08	0.00	0.00	-0.17	-0.02	-0.02	0.00	-0.01	0.01	-0.16
	0.00	1.00	1.00	0.00	0.02	0.00	0.43	0.16	0.02	0.00
Grim (PDB)	0.08	0.17	0.17	0.00	0.14	0.14	0.17	0.15	0.17	0.01
	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.83
Grim	-0.06	0.02	0.02	-0.14	0.00	0.00	0.03	0.01	0.03	-0.14
	0.00	0.02	0.02	0.00	1.00	0.96	0.01	0.32	0.00	0.00
IsoMIF	-0.06	0.02	0.02	-0.14	0.00	0.00	0.03	0.01	0.03	-0.14
	0.00	0.00	0.00	0.00	0.96	1.00	0.00	0.22	0.00	0.00
KRIPO	-0.09	0.00	0.00	-0.17	-0.03	-0.03	0.00	-0.01	0.00	-0.16
	0.00	0.43	0.43	0.00	0.01	0.00	1.00	0.07	0.00	0.00
PocketMatch	-0.07	0.01	0.01	-0.15	-0.01	-0.01	0.01	0.00	0.02	-0.15
	0.00	0.16	0.16	0.00	0.32	0.22	0.07	1.00	0.02	0.00
ProBiS	-0.09	-0.01	-0.01	-0.17	-0.03	-0.03	0.00	-0.02	0.00	-0.17
	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.02	NA	0.00
RAPMAD	0.08	0.16	0.16	-0.01	0.14	0.14	0.16	0.15	0.17	0.00
	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.00	1.00
VolSite/ Shaper (PDB)	-0.09	-0.01	-0.01	-0.17	-0.03	-0.03	0.00	-0.02	0.00	-0.17
	0.00	0.07	0.07	0.00	0.00	0.00	0.08	0.03	0.04	0.00
VolSite/ Shaper	-0.08	0.01	0.01	-0.16	-0.02	-0.02	0.01	-0.01	0.01	-0.15
	0.00	0.39	0.39	0.00	0.11	0.05	0.18	0.56	0.04	0.00
Shaper (PDB)	-0.02	0.06	0.06	-0.10	0.04	0.04	0.06	0.05	0.07	-0.10
	0.32	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00
Shaper	-0.02	0.06	0.06	-0.10	0.04	0.04	0.07	0.05	0.07	-0.10
	0.35	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00
SiteAlign	-0.09	-0.01	-0.01	-0.17	-0.03	-0.03	0.00	-0.02	0.00	-0.17
	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.02	0.22	0.00
SiteEngine	-0.09	-0.01	-0.01	-0.17	-0.03	-0.03	0.00	-0.02	0.00	-0.17
	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.02	NA	0.00
SiteHopper	-0.03	0.05	0.05	-0.11	0.03	0.03	0.06	0.04	0.06	-0.11
	0.20	0.00	0.00	0.00	0.10	0.08	0.00	0.01	0.00	0.00
SMAP	-0.09	-0.01	-0.01	-0.17	-0.03	-0.03	0.00	-0.02	0.00	-0.17
	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.02	NA	0.00
TIFP (PDB)	0.17	0.26	0.26	0.09	0.23	0.23	0.26	0.24	0.26	0.10
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TIFP	0.01	0.09	0.09	-0.07	0.07	0.07	0.10	0.08	0.10	-0.07
	0.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TM-align	-0.09	-0.01	-0.01	-0.17	-0.03	-0.03	0.00	-0.02	0.00	-0.17
	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.02	NA	0.00

Table A2.9 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 1.2. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	1.00 - 1.00	0.98 - 1.00	0.91 - 0.96	0.90 - 0.96	1.00 - 1.00	1.00 - 1.00	0.91 - 0.97	1.00 - 1.00	0.70 - 0.78	0.87 - 0.92	1.00 - 1.00
Cavbase	0.09	0.08	0.02	0.02	0.09	0.09	0.03	0.09	-0.17	-0.01	0.09
	0.00	0.00	0.32	0.35	0.00	0.00	0.20	0.00	0.00	0.62	0.00
FuzCav (PDB)	0.01	-0.01	-0.06	-0.06	0.01	0.01	-0.05	0.01	-0.26	-0.09	0.01
	0.07	0.39	0.00	0.00	0.02	0.02	0.00	0.02	0.00	0.00	0.02
FuzCav	0.01	-0.01	-0.06	-0.06	0.01	0.01	-0.05	0.01	-0.26	-0.09	0.01
	0.07	0.39	0.00	0.00	0.02	0.02	0.00	0.02	0.00	0.00	0.02
Grim (PDB)	0.17	0.16	0.10	0.10	0.17	0.17	0.11	0.17	-0.09	0.07	0.17
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim	0.03	0.02	-0.04	-0.04	0.03	0.03	-0.03	0.03	-0.23	-0.07	0.03
	0.00	0.11	0.02	0.02	0.00	0.00	0.10	0.00	0.00	0.00	0.00
IsoMIF	0.03	0.02	-0.04	-0.04	0.03	0.03	-0.03	0.03	-0.23	-0.07	0.03
	0.00	0.05	0.01	0.01	0.00	0.00	0.08	0.00	0.00	0.00	0.00
KRIPO	0.00	-0.01	-0.06	-0.07	0.00	0.00	-0.06	0.00	-0.26	-0.10	0.00
	0.08	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pocket-Match	0.02	0.01	-0.05	-0.05	0.02	0.02	-0.04	0.02	-0.24	-0.08	0.02
	0.03	0.56	0.00	0.00	0.02	0.02	0.01	0.02	0.00	0.00	0.02
ProBiS	0.00	-0.01	-0.07	-0.07	0.00	0.00	-0.06	0.00	-0.26	-0.10	0.00
	0.04	0.04	0.00	0.00	0.22	NA	0.00	NA	0.00	0.00	NA
RAPMAD	0.17	0.15	0.10	0.10	0.17	0.17	0.11	0.17	-0.10	0.07	0.17
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper (PDB)	0.00	-0.01	-0.07	-0.07	0.00	0.00	-0.06	0.00	-0.26	-0.10	0.00
	1.00	0.07	0.00	0.00	0.04	0.04	0.00	0.04	0.00	0.00	0.04
VolSite/ Shaper	0.01	0.00	-0.06	-0.06	0.01	0.01	-0.05	0.01	-0.25	-0.09	0.01
	0.07	1.00	0.00	0.00	0.04	0.04	0.00	0.04	0.00	0.00	0.04
Shaper (PDB)	0.07	0.06	0.00	0.00	0.07	0.07	0.01	0.07	-0.19	-0.03	0.07
	0.00	0.00	1.00	0.95	0.00	0.00	0.68	0.00	0.00	0.08	0.00
Shaper	0.07	0.06	0.00	0.00	0.07	0.07	0.01	0.07	-0.19	-0.03	0.07
	0.00	0.00	0.95	1.00	0.00	0.00	0.65	0.00	0.00	0.09	0.00
SiteAlign	0.00	-0.01	-0.07	-0.07	0.00	0.00	-0.06	0.00	-0.26	-0.10	0.00
	0.04	0.04	0.00	0.00	1.00	0.22	0.00	0.22	0.00	0.00	0.22
SiteEngine	0.00	-0.01	-0.07	-0.07	0.00	0.00	-0.06	0.00	-0.26	-0.10	0.00
	0.04	0.04	0.00	0.00	0.22	NA	0.00	NA	0.00	0.00	NA
SiteHopper	0.06	0.05	-0.01	-0.01	0.06	0.06	0.00	0.06	-0.20	-0.04	0.06
	0.00	0.00	0.68	0.65	0.00	0.00	1.00	0.00	0.00	0.04	0.00
SMAP	0.00	-0.01	-0.07	-0.07	0.00	0.00	-0.06	0.00	-0.26	-0.10	0.00
	0.04	0.04	0.00	0.00	0.22	NA	0.00	NA	0.00	0.00	NA
TIFP (PDB)	0.26	0.25	0.19	0.19	0.26	0.26	0.20	0.26	0.00	0.16	0.26
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
TIFP	0.10	0.09	0.03	0.03	0.10	0.10	0.04	0.10	-0.16	0.00	0.10
	0.00	0.00	0.08	0.09	0.00	0.00	0.04	0.00	0.00	1.00	0.00
TM-align	0.00	-0.01	-0.07	-0.07	0.00	0.00	-0.06	0.00	-0.26	-0.10	0.00
	0.04	0.04	0.00	0.00	0.22	NA	0.00	NA	0.00	0.00	NA

Table A2.10. AUC and EFs of different binding site comparison methods for dataset 2.

Method	AUC	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	0.87	13.97	14.00	14.00	14.00	13.99	13.14	12.26
FuzCav	0.99	13.97	14.00	14.00	14.00	14.00	14.00	13.96
FuzCav (PDB)	0.98	13.97	14.00	14.00	14.00	14.00	13.99	13.94
Grim	0.92	13.97	14.00	14.00	14.00	13.99	13.83	12.58
Grim (PDB)	0.85	13.97	13.82	13.91	13.51	13.36	12.55	11.26
IsoMIF	0.70	13.97	13.92	13.90	13.28	11.81	10.24	8.92
KRIPO	0.96	13.97	14.00	14.00	14.00	13.94	13.61	12.95
PocketMatch	0.96	13.97	14.00	14.00	14.00	14.00	13.94	13.50
ProBiS	1.00	13.97	13.97	13.99	13.99	14.00	14.00	14.00
RAPMAD	0.82	13.97	14.00	13.92	12.93	11.71	10.75	9.88
Shaper	0.93	13.97	14.00	14.00	13.96	13.84	13.16	12.05
Shaper (PDB)	0.93	13.97	14.00	14.00	13.96	13.84	13.22	12.16
VolSite/ Shaper	0.78	13.97	14.00	14.00	14.00	13.96	13.61	12.31
VolSite/ Shaper (PDB)	0.76	13.97	14.00	14.00	14.00	13.94	13.24	11.64
SiteAlign	1.00	13.97	14.00	14.00	14.00	14.00	14.00	14.00
SiteEngine	1.00	13.97	14.00	14.00	14.00	14.00	14.00	14.00
SiteHopper	1.00	13.97	14.00	14.00	14.00	14.00	14.00	14.00
SMAP	1.00	13.97	14.00	14.00	14.00	14.00	14.00	14.00
TIFP	0.91	13.97	13.64	13.05	11.74	10.80	9.56	8.38
TIFP (PDB)	0.78	5.95	12.39	12.63	11.24	9.85	8.80	7.77
TM-align	1.00	13.97	14.00	14.00	14.00	14.00	14.00	14.00

Table A2.11. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 2. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.86 - 0.88	0.98 - 0.99	0.98 - 0.99	0.84 - 0.85	0.91 - 0.92	0.69 - 0.70	0.96 - 0.96	0.96 - 0.96	1.00 - 1.00	0.81 - 0.83
Cavbase	0.00	0.11	0.11	-0.02	0.04	-0.18	0.09	0.09	0.13	-0.05
	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav (PDB)	-0.11	0.00	0.00	-0.14	-0.07	-0.29	-0.02	-0.03	0.01	-0.17
	0.00	1.00	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav	-0.11	0.00	0.00	-0.14	-0.07	-0.29	-0.02	-0.03	0.01	-0.17
	0.00	0.44	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim (PDB)	0.02	0.14	0.14	0.00	0.07	-0.15	0.11	0.11	0.15	-0.03
	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim	-0.04	0.07	0.07	-0.07	0.00	-0.22	0.05	0.04	0.08	-0.10
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
IsoMIF	0.18	0.29	0.29	0.15	0.22	0.00	0.27	0.26	0.30	0.12
	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
KRIPO	-0.09	0.02	0.02	-0.11	-0.05	-0.27	0.00	0.00	0.04	-0.14
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.18	0.00	0.00
PocketMatch	-0.09	0.03	0.03	-0.11	-0.04	-0.26	0.00	0.00	0.04	-0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.18	1.00	0.00	0.00
ProBiS	-0.13	-0.01	-0.01	-0.15	-0.08	-0.30	-0.04	-0.04	0.00	-0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
RAPMAD	0.05	0.17	0.17	0.03	0.10	-0.12	0.14	0.14	0.18	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
VolSite/ Shaper (PDB)	0.11	0.23	0.23	0.09	0.16	-0.06	0.20	0.20	0.24	0.06
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper	0.09	0.20	0.20	0.06	0.13	-0.09	0.18	0.17	0.21	0.03
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper (PDB)	-0.06	0.05	0.06	-0.08	-0.01	-0.23	0.03	0.03	0.07	-0.11
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper	-0.06	0.06	0.06	-0.08	-0.01	-0.23	0.03	0.03	0.07	-0.11
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteAlign	-0.13	-0.01	-0.01	-0.15	-0.08	-0.30	-0.04	-0.04	0.00	-0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00
SiteEngine	-0.13	-0.01	-0.01	-0.15	-0.08	-0.30	-0.04	-0.04	0.00	-0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
SiteHopper	-0.13	-0.02	-0.01	-0.15	-0.08	-0.30	-0.04	-0.04	0.00	-0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SMAP	-0.13	-0.02	-0.01	-0.15	-0.08	-0.30	-0.04	-0.04	0.00	-0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TIFP (PDB)	0.09	0.20	0.20	0.07	0.13	-0.09	0.18	0.18	0.22	0.04
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TIFP	-0.04	0.08	0.08	-0.06	0.01	-0.21	0.05	0.05	0.09	-0.09
	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
TM-align	-0.13	-0.02	-0.01	-0.15	-0.08	-0.30	-0.04	-0.04	0.00	-0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A2.11 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 2. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.75 - 0.76	0.78 - 0.79	0.93 - 0.93	0.92 - 0.93	1.00 - 1.00	1.00 - 1.00	1.00 - 1.00	1.00 - 1.00	0.78 - 0.79	0.90 - 0.91	1.00 - 1.00
Cavbase	-0.11	-0.09	0.06	0.06	0.13	0.13	0.13	0.13	-0.09	0.04	0.13
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav (PDB)	-0.23	-0.20	-0.05	-0.06	0.01	0.01	0.02	0.02	-0.20	-0.08	0.02
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav	-0.23	-0.20	-0.06	-0.06	0.01	0.01	0.01	0.01	-0.20	-0.08	0.01
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim (PDB)	-0.09	-0.06	0.08	0.08	0.15	0.15	0.15	0.15	-0.07	0.06	0.15
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim	-0.16	-0.13	0.01	0.01	0.08	0.08	0.08	0.08	-0.13	-0.01	0.08
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
IsoMIF	0.06	0.09	0.23	0.23	0.30	0.30	0.30	0.30	0.09	0.21	0.30
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KRIPO	-0.20	-0.18	-0.03	-0.03	0.04	0.04	0.04	0.04	-0.18	-0.05	0.04
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pocket-Match	-0.20	-0.17	-0.03	-0.03	0.04	0.04	0.04	0.04	-0.18	-0.05	0.04
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ProBiS	-0.24	-0.21	-0.07	-0.07	0.00	0.00	0.00	0.00	-0.22	-0.09	0.00
	0.00	0.00	0.00	0.00	0.29	0.10	0.00	0.00	0.00	0.00	0.00
RAPMAD	-0.06	-0.03	0.11	0.11	0.18	0.18	0.18	0.18	-0.04	0.09	0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper (PDB)	0.00	0.03	0.17	0.17	0.24	0.24	0.24	0.24	0.03	0.15	0.24
	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper	-0.03	0.00	0.15	0.14	0.21	0.21	0.22	0.22	0.00	0.12	0.22
	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.70	0.00	0.00
Shaper (PDB)	-0.17	-0.15	0.00	0.00	0.07	0.07	0.07	0.07	-0.15	-0.02	0.07
	0.00	0.00	1.00	0.76	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper	-0.17	-0.14	0.00	0.00	0.07	0.07	0.07	0.07	-0.15	-0.02	0.07
	0.00	0.00	0.76	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteAlign	-0.24	-0.21	-0.07	-0.07	0.00	0.00	0.00	0.00	-0.22	-0.09	0.00
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteEngine	-0.24	-0.21	-0.07	-0.07	0.00	0.00	0.00	0.00	-0.22	-0.09	0.00
	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
SiteHopper	-0.24	-0.22	-0.07	-0.07	0.00	0.00	0.00	0.00	-0.22	-0.09	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.07	0.00	0.00	0.00
SMAP	-0.24	-0.22	-0.07	-0.07	0.00	0.00	0.00	0.00	-0.22	-0.09	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.07	1.00	0.00	0.00	0.00
TIFP (PDB)	-0.03	0.00	0.15	0.15	0.22	0.22	0.22	0.22	0.00	0.13	0.22
	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
TIFP	-0.15	-0.12	0.02	0.02	0.09	0.09	0.09	0.09	-0.13	0.00	0.09
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
TM-align	-0.24	-0.22	-0.07	-0.07	0.00	0.00	0.00	0.00	-0.22	-0.09	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table A2.12. AUC and EFs of different binding site comparison methods for dataset 3.

Method	AUC	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	0.65	1.41	1.65	1.36	1.05	1.13	0.97	1.03
FuzCav	0.69	1.94	2.00	2.00	1.94	1.87	1.80	1.78
FuzCav (PDB)	0.69	1.94	2.00	2.00	1.94	1.87	1.80	1.78
Grim	0.55	1.64	1.30	1.23	1.34	1.42	1.38	1.38
Grim (PDB)	0.57	1.94	1.97	1.83	1.63	1.53	1.38	1.30
IsoMIF	0.59	1.94	2.00	2.00	1.58	1.51	1.49	1.47
KRIPO	0.60	1.94	2.00	2.00	1.64	1.57	1.54	1.50
PocketMatch	0.59	0.00	0.00	0.89	1.45	1.63	1.64	1.58
ProBiS	0.47	1.71	1.53	1.21	1.00	1.00	1.03	0.99
RAPMAD	0.61	1.94	2.00	2.00	1.72	1.70	1.70	1.69
Shaper	0.71	1.94	2.00	2.00	1.63	1.61	1.64	1.66
Shaper (PDB)	0.71	1.94	2.00	2.00	1.62	1.61	1.64	1.64
VolSite/ Shaper	0.68	1.94	2.00	2.00	1.68	1.66	1.66	1.68
VolSite/ Shaper (PDB)	0.68	1.94	2.00	2.00	1.68	1.66	1.65	1.68
SiteAlign	0.85	1.94	2.00	2.00	2.00	2.00	2.00	2.00
SiteEngine	0.82	1.94	1.98	1.95	1.83	1.61	1.55	1.53
SiteHopper	0.75	1.94	2.00	2.00	1.86	1.79	1.79	1.80
SMAP	0.76	1.19	1.82	1.88	1.93	1.94	1.95	1.95
TIFP	0.66	1.71	1.95	1.97	1.62	1.54	1.54	1.51
TIFP (PDB)	0.56	0.00	1.09	1.54	1.58	1.49	1.44	1.34
TM-align	0.49	0.00	0.00	0.00	0.79	0.94	0.84	0.78

Table A2.13. AUC and EFs of different binding site comparison methods for dataset 4.

Method	AUC	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	0.64	1.71	1.62	1.63	1.08	1.17	1.05	0.99
FuzCav	0.58	1.49	1.91	1.95	1.79	1.56	1.50	1.50
FuzCav (PDB)	0.58	1.49	1.91	1.95	1.79	1.56	1.50	1.50
Grim	0.56	1.64	1.33	1.36	1.43	1.43	1.39	1.38
Grim (PDB)	0.56	1.94	1.77	1.82	1.59	1.58	1.38	1.30
IsoMIF	0.59	1.94	2.00	2.00	1.69	1.63	1.60	1.58
KRIPO	0.61	1.94	2.00	2.00	1.58	1.42	1.36	1.32
PocketMatch	0.57	0.00	0.00	0.83	1.42	1.61	1.64	1.58
ProBiS	0.46	1.64	1.71	1.27	0.89	0.95	0.99	0.95
RAPMAD	0.63	1.94	2.00	2.00	1.74	1.72	1.72	1.72
Shaper	0.76	1.94	2.00	2.00	1.75	1.70	1.71	1.71
Shaper (PDB)	0.76	1.94	2.00	2.00	1.75	1.71	1.70	1.71
VolSite/ Shaper	0.76	1.94	2.00	2.00	1.83	1.78	1.76	1.77
VolSite/ Shaper (PDB)	0.76	1.94	2.00	2.00	1.82	1.78	1.76	1.77
SiteAlign	0.80	1.94	2.00	2.00	1.96	1.94	1.88	1.88
SiteEngine	0.79	1.94	1.98	1.98	1.99	1.93	1.87	1.82
SiteHopper	0.75	1.94	2.00	2.00	1.92	1.88	1.86	1.84
SMAP	0.65	1.19	1.85	1.92	1.94	1.90	1.87	1.82
TIFP	0.66	0.82	1.77	1.88	1.66	1.58	1.58	1.54
TIFP (PDB)	0.57	0.00	1.00	1.50	1.55	1.48	1.45	1.39
TM-align	0.49	0.00	0.00	0.00	0.79	0.94	0.84	0.78

Table A2.14. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 3. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.64 - 0.65	0.68 - 0.69	0.68 - 0.69	0.56 - 0.57	0.54 - 0.55	0.59 - 0.60	0.59 - 0.61	0.58 - 0.59	0.46 - 0.48	0.60 - 0.61
Cavbase	0.00	0.04	0.04	-0.08	-0.10	-0.05	-0.05	-0.06	-0.18	-0.04
	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav (PDB)	-0.04	0.00	0.00	-0.12	-0.14	-0.09	-0.09	-0.10	-0.22	-0.08
	0.00	1.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav	-0.04	0.00	0.00	-0.12	-0.14	-0.09	-0.09	-0.10	-0.22	-0.08
	0.00	0.99	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim (PDB)	0.08	0.12	0.12	0.00	-0.02	0.03	0.03	0.02	-0.10	0.04
	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim	0.10	0.14	0.14	0.02	0.00	0.05	0.05	0.04	-0.07	0.06
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
IsoMIF	0.05	0.09	0.09	-0.03	-0.05	0.00	0.01	-0.01	-0.12	0.01
	0.00	0.00	0.00	0.00	0.00	1.00	0.19	0.14	0.00	0.00
KRIPO	0.05	0.09	0.09	-0.03	-0.05	-0.01	0.00	-0.01	-0.13	0.01
	0.00	0.00	0.00	0.00	0.00	0.19	1.00	0.01	0.00	0.12
PocketMatch	0.06	0.10	0.10	-0.02	-0.04	0.01	0.01	0.00	-0.11	0.02
	0.00	0.00	0.00	0.00	0.00	0.14	0.01	1.00	0.00	0.00
ProBiS	0.18	0.22	0.22	0.10	0.07	0.12	0.13	0.11	0.00	0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
RAPMAD	0.04	0.08	0.08	-0.04	-0.06	-0.01	-0.01	-0.02	-0.14	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	1.00
VolSite/ Shaper (PDB)	-0.03	0.01	0.01	-0.11	-0.13	-0.08	-0.08	-0.09	-0.20	-0.07
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper	-0.03	0.01	0.01	-0.11	-0.13	-0.08	-0.08	-0.09	-0.20	-0.07
	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper (PDB)	-0.06	-0.02	-0.02	-0.14	-0.16	-0.12	-0.11	-0.12	-0.24	-0.10
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper	-0.06	-0.02	-0.02	-0.14	-0.17	-0.12	-0.11	-0.13	-0.24	-0.10
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteAlign	-0.20	-0.16	-0.16	-0.28	-0.30	-0.26	-0.25	-0.26	-0.38	-0.24
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteEngine	-0.17	-0.13	-0.13	-0.25	-0.27	-0.23	-0.22	-0.23	-0.35	-0.21
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteHopper	-0.10	-0.06	-0.06	-0.18	-0.20	-0.16	-0.15	-0.16	-0.28	-0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SMAP	-0.11	-0.07	-0.07	-0.19	-0.21	-0.16	-0.16	-0.17	-0.29	-0.15
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TIFP (PDB)	0.09	0.13	0.13	0.01	-0.01	0.04	0.04	0.03	-0.09	0.05
	0.00	0.00	0.00	0.02	0.04	0.00	0.00	0.00	0.00	0.00
TIFP	-0.01	0.03	0.03	-0.09	-0.11	-0.06	-0.06	-0.07	-0.18	-0.05
	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TM-align	0.16	0.20	0.20	0.08	0.06	0.10	0.11	0.09	-0.02	0.12
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2.14 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 3. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.67 - 0.68	0.67 - 0.68	0.70 - 0.72	0.71 - 0.72	0.85 - 0.86	0.81 - 0.83	0.74 - 0.76	0.75 - 0.76	0.55 - 0.56	0.65 - 0.66	0.48 - 0.50
Cavbase	0.03	0.03	0.06	0.06	0.20	0.17	0.10	0.11	-0.09	0.01	-0.16
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00
FuzCav (PDB)	-0.01	-0.01	0.02	0.02	0.16	0.13	0.06	0.07	-0.13	-0.03	-0.20
	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav	-0.01	-0.01	0.02	0.02	0.16	0.13	0.06	0.07	-0.13	-0.03	-0.20
	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim (PDB)	0.11	0.11	0.14	0.14	0.28	0.25	0.18	0.19	-0.01	0.09	-0.08
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
Grim	0.13	0.13	0.16	0.17	0.30	0.27	0.20	0.21	0.01	0.11	-0.06
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00
IsoMIF	0.08	0.08	0.12	0.12	0.26	0.23	0.16	0.16	-0.04	0.06	-0.10
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KRIPO	0.08	0.08	0.11	0.11	0.25	0.22	0.15	0.16	-0.04	0.06	-0.11
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pocket-Match	0.09	0.09	0.12	0.13	0.26	0.23	0.16	0.17	-0.03	0.07	-0.09
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ProBiS	0.20	0.20	0.24	0.24	0.38	0.35	0.28	0.29	0.09	0.18	0.02
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPMAD	0.07	0.07	0.10	0.10	0.24	0.21	0.14	0.15	-0.05	0.05	-0.12
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper (PDB)	0.00	0.00	0.03	0.04	0.18	0.14	0.07	0.08	-0.12	-0.02	-0.18
	1.00	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper	0.00	0.00	0.03	0.04	0.17	0.14	0.07	0.08	-0.12	-0.02	-0.18
	0.90	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper (PDB)	-0.03	-0.03	0.00	0.00	0.14	0.11	0.04	0.05	-0.15	-0.05	-0.22
	0.00	0.00	1.00	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper	-0.04	-0.04	0.00	0.00	0.14	0.11	0.04	0.05	-0.16	-0.06	-0.22
	0.00	0.00	0.55	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteAlign	-0.18	-0.17	-0.14	-0.14	0.00	-0.03	-0.10	-0.09	-0.29	-0.19	-0.36
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteEngine	-0.14	-0.14	-0.11	-0.11	0.03	0.00	-0.07	-0.06	-0.26	-0.16	-0.33
	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
SiteHopper	-0.07	-0.07	-0.04	-0.04	0.10	0.07	0.00	0.01	-0.19	-0.09	-0.26
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.06	0.00	0.00	0.00
SMAP	-0.08	-0.08	-0.05	-0.05	0.09	0.06	-0.01	0.00	-0.20	-0.10	-0.27
	0.00	0.00	0.00	0.00	0.00	0.00	0.06	1.00	0.00	0.00	0.00
TIFP (PDB)	0.12	0.12	0.15	0.16	0.29	0.26	0.19	0.20	0.00	0.10	-0.07
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
TIFP	0.02	0.02	0.05	0.06	0.19	0.16	0.09	0.10	-0.10	0.00	-0.17
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
TM-align	0.18	0.18	0.22	0.22	0.36	0.33	0.26	0.27	0.07	0.17	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table A2.15. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 4. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.64 - 0.65	0.58 - 0.59	0.58 - 0.59	0.55 - 0.57	0.55 - 0.56	0.58 - 0.60	0.60 - 0.61	0.56 - 0.57	0.46 - 0.47	0.63 - 0.64
Cavbase	0.00	-0.06	-0.06	-0.08	-0.09	-0.05	-0.04	-0.08	-0.18	-0.01
	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
FuzCav (PDB)	0.06	0.00	0.00	-0.02	-0.03	0.01	0.02	-0.02	-0.12	0.05
	0.00	1.00	0.93	0.00	0.00	0.31	0.00	0.00	0.00	0.00
FuzCav	0.06	0.00	0.00	-0.02	-0.03	0.01	0.02	-0.02	-0.12	0.05
	0.00	0.93	1.00	0.00	0.00	0.27	0.00	0.00	0.00	0.00
Grim (PDB)	0.08	0.02	0.02	0.00	0.00	0.03	0.05	0.01	-0.10	0.07
	0.00	0.00	0.00	1.00	0.39	0.00	0.00	0.17	0.00	0.00
Grim	0.09	0.03	0.03	0.00	0.00	0.03	0.05	0.01	-0.09	0.08
	0.00	0.00	0.00	0.39	1.00	0.00	0.00	0.03	0.00	0.00
IsoMIF	0.05	-0.01	-0.01	-0.03	-0.03	0.00	0.02	-0.02	-0.13	0.04
	0.00	0.31	0.27	0.00	0.00	1.00	0.00	0.00	0.00	0.00
KRIPO	0.04	-0.02	-0.02	-0.05	-0.05	-0.02	0.00	-0.04	-0.14	0.03
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
PocketMatch	0.08	0.02	0.02	-0.01	-0.01	0.02	0.04	0.00	-0.10	0.07
	0.00	0.00	0.00	0.17	0.03	0.00	0.00	1.00	0.00	0.00
ProBiS	0.18	0.12	0.12	0.10	0.09	0.13	0.14	0.10	0.00	0.17
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
RAPMAD	0.01	-0.05	-0.05	-0.07	-0.08	-0.04	-0.03	-0.07	-0.17	0.00
	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
VolSite/ Shaper (PDB)	-0.12	-0.17	-0.17	-0.20	-0.20	-0.17	-0.15	-0.19	-0.29	-0.13
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper	-0.11	-0.17	-0.17	-0.20	-0.20	-0.17	-0.15	-0.19	-0.29	-0.12
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper (PDB)	-0.12	-0.18	-0.18	-0.20	-0.20	-0.17	-0.15	-0.19	-0.30	-0.13
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper	-0.12	-0.18	-0.18	-0.20	-0.20	-0.17	-0.15	-0.19	-0.30	-0.13
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteAlign	-0.16	-0.22	-0.22	-0.24	-0.25	-0.21	-0.20	-0.24	-0.34	-0.17
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteEngine	-0.14	-0.20	-0.20	-0.22	-0.23	-0.20	-0.18	-0.22	-0.32	-0.15
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteHopper	-0.11	-0.17	-0.17	-0.19	-0.19	-0.16	-0.14	-0.18	-0.29	-0.12
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SMAP	-0.01	-0.07	-0.07	-0.09	-0.10	-0.06	-0.05	-0.08	-0.19	-0.02
	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TIFP (PDB)	0.08	0.02	0.02	-0.01	-0.01	0.02	0.04	0.00	-0.10	0.07
	0.00	0.00	0.00	0.29	0.06	0.00	0.00	0.76	0.00	0.00
TIFP	-0.02	-0.08	-0.08	-0.10	-0.11	-0.07	-0.06	-0.10	-0.20	-0.03
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TM-align	0.15	0.09	0.09	0.07	0.07	0.10	0.12	0.08	-0.03	0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A2.15 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 4. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.75 - 0.76	0.75 - 0.76	0.76 - 0.77	0.75 - 0.77	0.80 - 0.81	0.78 - 0.79	0.74 - 0.76	0.65 - 0.66	0.56 - 0.57	0.66 - 0.67	0.48 - 0.50
Cavbase	0.12	0.11	0.12	0.12	0.16	0.14	0.11	0.01	-0.08	0.02	-0.15
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00
FuzCav (PDB)	0.17	0.17	0.18	0.18	0.22	0.20	0.17	0.07	-0.02	0.08	-0.09
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav	0.17	0.17	0.18	0.18	0.22	0.20	0.17	0.07	-0.02	0.08	-0.09
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim (PDB)	0.20	0.20	0.20	0.20	0.24	0.22	0.19	0.09	0.01	0.10	-0.07
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00
Grim	0.20	0.20	0.20	0.20	0.25	0.23	0.19	0.10	0.01	0.11	-0.07
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00
IsoMIF	0.17	0.17	0.17	0.17	0.21	0.20	0.16	0.06	-0.02	0.07	-0.10
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KRIPO	0.15	0.15	0.15	0.15	0.20	0.18	0.14	0.05	-0.04	0.06	-0.12
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pocket-Match	0.19	0.19	0.19	0.19	0.24	0.22	0.18	0.08	0.00	0.10	-0.08
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.76	0.00	0.00
ProBiS	0.29	0.29	0.30	0.30	0.34	0.32	0.29	0.19	0.10	0.20	0.03
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPMAD	0.13	0.12	0.13	0.13	0.17	0.15	0.12	0.02	-0.07	0.03	-0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper (PDB)	0.00	0.00	0.00	0.00	0.05	0.03	-0.01	-0.11	-0.19	-0.09	-0.27
	1.00	0.94	0.48	0.54	0.00	0.00	0.05	0.00	0.00	0.00	0.00
VolSite/ Shaper	0.00	0.00	0.00	0.00	0.05	0.03	-0.01	-0.11	-0.19	-0.09	-0.27
	0.94	1.00	0.43	0.50	0.00	0.00	0.06	0.00	0.00	0.00	0.00
Shaper (PDB)	0.00	0.00	0.00	0.00	0.04	0.02	-0.01	-0.11	-0.20	-0.10	-0.27
	0.48	0.43	1.00	0.92	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Shaper	0.00	0.00	0.00	0.00	0.04	0.03	-0.01	-0.11	-0.19	-0.10	-0.27
	0.54	0.50	0.92	1.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
SiteAlign	-0.05	-0.05	-0.04	-0.04	0.00	-0.02	-0.05	-0.15	-0.24	-0.14	-0.31
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteEngine	-0.03	-0.03	-0.02	-0.03	0.02	0.00	-0.04	-0.13	-0.22	-0.12	-0.29
	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
SiteHopper	0.01	0.01	0.01	0.01	0.05	0.04	0.00	-0.10	-0.18	-0.09	-0.26
	0.05	0.06	0.01	0.01	0.00	0.00	1.00	0.00	0.00	0.00	0.00
SMAP	0.11	0.11	0.11	0.11	0.15	0.13	0.10	0.00	-0.09	0.01	-0.16
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.01	0.00
TIFP (PDB)	0.19	0.19	0.20	0.19	0.24	0.22	0.18	0.09	0.00	0.10	-0.07
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
TIFP	0.09	0.09	0.10	0.10	0.14	0.12	0.09	-0.01	-0.10	0.00	-0.17
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	1.00	0.00
TM-align	0.27	0.27	0.27	0.27	0.31	0.29	0.26	0.16	0.07	0.17	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table A2.16. AUC and EFs of different binding site comparison methods for dataset 5.

Method	AUC	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	0.58	4.16	2.51	1.72	1.30	1.15	1.07	1.03
FuzCav	0.57	0.00	0.50	0.63	0.90	1.06	1.14	1.17
FuzCav (PDB)	0.57	0.00	0.50	0.70	0.96	1.11	1.18	1.18
Grim	0.73	3.46	4.30	4.39	4.26	4.12	3.96	3.81
Grim (PDB)	0.64	0.00	1.49	1.20	1.41	1.66	1.77	1.82
IsoMIF	0.97	4.63	4.60	4.62	4.62	4.62	4.62	4.61
KRIPO	0.93	4.63	4.63	4.58	4.50	4.34	4.26	4.15
PocketMatch	0.79	4.56	4.47	4.37	4.23	4.02	3.85	3.68
ProBiS	0.52	4.63	4.60	4.28	3.55	3.21	2.90	2.50
RAPMAD	0.60	2.48	2.74	2.50	2.27	2.04	1.93	1.87
Shaper	0.68	4.24	3.35	3.07	2.79	2.59	2.45	2.36
Shaper (PDB)	0.68	4.37	3.26	3.05	2.77	2.59	2.39	2.30
VolSite/ Shaper	0.68	3.00	3.27	3.12	2.83	2.67	2.53	2.45
VolSite/ Shaper (PDB)	0.70	2.67	2.96	2.86	2.73	2.62	2.54	2.49
SiteAlign	0.46	0.98	0.48	0.44	0.47	0.49	0.51	0.50
SiteEngine	0.56	2.46	1.66	1.40	1.18	1.10	1.06	1.04
SiteHopper	0.97	4.63	4.65	4.66	4.66	4.66	4.66	4.66
SMAP	0.56	4.63	4.65	4.24	3.11	2.44	2.07	1.82
TIFP	0.57	4.17	3.60	3.09	2.58	2.22	2.06	1.92
TIFP (PDB)	0.56	1.89	1.73	1.88	1.81	1.69	1.73	1.68
TM-align	0.79	4.63	4.65	4.66	4.65	4.62	4.55	4.46

Table A2.17. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 5. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.58 - 0.59	0.57 - 0.58	0.57 - 0.58	0.64 - 0.65	0.73- 0.74	0.97- 0.97	0.93- 0.93	0.79- 0.80	0.52- 0.53	0.60- 0.61
Cavbase	0.00	-0.01	-0.01	0.06	0.15	0.39	0.34	0.21	-0.06	0.02
	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav (PDB)	0.01	0.00	0.00	0.07	0.16	0.40	0.36	0.22	-0.05	0.03
	0.00	1.00	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav	0.01	0.00	0.00	0.07	0.16	0.40	0.35	0.22	-0.05	0.03
	0.00	0.37	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim (PDB)	-0.06	-0.07	-0.07	0.00	0.09	0.33	0.28	0.15	-0.12	-0.04
	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim	-0.15	-0.16	-0.16	-0.09	0.00	0.24	0.19	0.06	-0.21	-0.13
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
IsoMIF	-0.39	-0.40	-0.40	-0.33	-0.24	0.00	-0.05	-0.18	-0.45	-0.37
	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
KRIPO	-0.34	-0.36	-0.35	-0.28	-0.19	0.05	0.00	-0.14	-0.40	-0.32
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
PocketMatch	-0.21	-0.22	-0.22	-0.15	-0.06	0.18	0.14	0.00	-0.27	-0.19
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
ProBiS	0.06	0.05	0.05	0.12	0.21	0.45	0.40	0.27	0.00	0.08
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
RAPMAD	-0.02	-0.03	-0.03	0.04	0.13	0.37	0.32	0.19	-0.08	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
VolSite/ Shaper (PDB)	-0.11	-0.13	-0.13	-0.06	0.03	0.27	0.23	0.09	-0.18	-0.10
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper	-0.09	-0.11	-0.11	-0.04	0.05	0.29	0.25	0.11	-0.16	-0.08
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper (PDB)	-0.09	-0.11	-0.10	-0.03	0.06	0.30	0.25	0.11	-0.15	-0.07
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper	-0.09	-0.11	-0.10	-0.03	0.06	0.30	0.25	0.11	-0.15	-0.07
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteAlign	0.13	0.11	0.11	0.18	0.27	0.51	0.47	0.33	0.07	0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteEngine	0.03	0.01	0.02	0.09	0.18	0.42	0.37	0.24	-0.03	0.05
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteHopper	-0.38	-0.40	-0.40	-0.33	-0.24	0.00	-0.04	-0.18	-0.45	-0.37
	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00
SMAP	0.03	0.01	0.02	0.09	0.18	0.42	0.37	0.23	-0.03	0.05
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TIFP (PDB)	0.02	0.01	0.01	0.08	0.17	0.41	0.37	0.23	-0.04	0.04
	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TIFP	0.02	0.00	0.01	0.08	0.17	0.41	0.36	0.22	-0.04	0.04
	0.00	0.47	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TM-align	-0.20	-0.22	-0.21	-0.14	-0.05	0.19	0.14	0.00	-0.26	-0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00

Table A2.17 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 5. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.69 - 0.70	0.67 - 0.68	0.67 - 0.68	0.67 - 0.68	0.45 - 0.46	0.55 - 0.56	0.97 - 0.97	0.55 - 0.56	0.56 - 0.57	0.56 - 0.57	0.78 - 0.79
Cavbase	0.11	0.09	0.09	0.09	-0.13	-0.03	0.38	-0.03	-0.02	-0.02	0.20
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav (PDB)	0.13	0.11	0.11	0.11	-0.11	-0.01	0.40	-0.01	-0.01	0.00	0.22
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.47	0.00
FuzCav	0.13	0.11	0.10	0.10	-0.11	-0.02	0.40	-0.02	-0.01	-0.01	0.21
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00
Grim (PDB)	0.06	0.04	0.03	0.03	-0.18	-0.09	0.33	-0.09	-0.08	-0.08	0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim	-0.03	-0.05	-0.06	-0.06	-0.27	-0.18	0.24	-0.18	-0.17	-0.17	0.05
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IsoMIF	-0.27	-0.29	-0.30	-0.30	-0.51	-0.42	0.00	-0.42	-0.41	-0.41	-0.19
	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00
KRIPO	-0.23	-0.25	-0.25	-0.25	-0.47	-0.37	0.04	-0.37	-0.37	-0.36	-0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pocket-Match	-0.09	-0.11	-0.11	-0.11	-0.33	-0.24	0.18	-0.23	-0.23	-0.22	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29
ProBiS	0.18	0.16	0.15	0.15	-0.07	0.03	0.45	0.03	0.04	0.04	0.26
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RAPMAD	0.10	0.08	0.07	0.07	-0.14	-0.05	0.37	-0.05	-0.04	-0.04	0.18
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper (PDB)	0.00	-0.02	-0.02	-0.02	-0.24	-0.14	0.27	-0.14	-0.14	-0.13	0.09
	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VolSite/ Shaper	0.02	0.00	0.00	0.00	-0.22	-0.12	0.29	-0.12	-0.12	-0.11	0.11
	0.00	1.00	0.39	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper (PDB)	0.02	0.00	0.00	0.00	-0.22	-0.12	0.29	-0.12	-0.12	-0.11	0.11
	0.00	0.39	1.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shaper	0.02	0.00	0.00	0.00	-0.22	-0.12	0.29	-0.12	-0.12	-0.11	0.11
	0.00	0.31	0.86	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteAlign	0.24	0.22	0.22	0.22	0.00	0.10	0.51	0.10	0.10	0.11	0.33
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
SiteEngine	0.14	0.12	0.12	0.12	-0.10	0.00	0.41	0.00	0.01	0.01	0.23
	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.53	0.16	0.00	0.00
SiteHopper	-0.27	-0.29	-0.29	-0.29	-0.51	-0.41	0.00	-0.41	-0.41	-0.40	-0.18
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
SMAP	0.14	0.12	0.12	0.12	-0.10	0.00	0.41	0.00	0.00	0.01	0.23
	0.00	0.00	0.00	0.00	0.00	0.53	0.00	1.00	0.39	0.01	0.00
TIFP (PDB)	0.14	0.12	0.12	0.12	-0.10	-0.01	0.41	0.00	0.00	0.01	0.23
	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.39	1.00	0.08	0.00
TIFP	0.13	0.11	0.11	0.11	-0.11	-0.01	0.40	-0.01	-0.01	0.00	0.22
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.08	1.00	0.00
TM-align	-0.09	-0.11	-0.11	-0.11	-0.33	-0.23	0.18	-0.23	-0.23	-0.22	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table A2.18. AUC and EFs of different binding site comparison methods for dataset 6.

Method	AUC	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	0.60	6.52	6.96	6.96	4.89	3.51	3.23	2.89
FuzCav	0.55	6.52	6.96	6.96	4.89	3.33	2.83	2.35
FuzCav (PDB)	0.56	6.52	6.96	6.96	4.89	3.33	2.88	2.43
Grim	0.69	6.52	6.96	6.20	5.87	5.36	4.73	4.39
Grim (PDB)	0.61	0.00	2.39	4.67	3.75	3.44	2.66	2.13
IsoMIF	0.75	6.52	6.96	6.96	5.43	5.18	5.14	4.76
KRIPO	0.76	6.52	6.96	6.96	5.87	5.07	4.62	4.22
PocketMatch	0.66	6.52	6.96	6.96	5.33	4.20	3.75	3.57
ProBiS	0.54	6.52	6.96	6.96	5.05	3.73	3.34	2.91
RAPMAD	0.55	6.52	6.96	6.96	4.89	3.48	2.72	2.35
Shaper	0.65	6.52	6.96	6.96	5.11	4.35	3.94	3.48
Shaper (PDB)	0.66	6.52	6.96	6.96	5.22	4.49	4.02	3.70
VolSite/ Shaper	0.56	6.52	6.96	6.52	4.13	3.15	2.74	2.41
VolSite/ Shaper (PDB)	0.57	6.52	6.96	6.09	4.02	3.12	2.77	2.50
SiteAlign	0.59	6.52	6.96	6.96	4.89	3.80	3.10	2.70
SiteEngine	0.64	6.52	6.96	6.96	5.38	4.35	3.64	3.17
SiteHopper	0.72	6.52	6.96	6.96	5.82	4.82	4.21	3.76
SMAP	0.62	6.52	6.96	6.96	5.11	3.91	3.07	2.78
TIFP	0.71	6.52	6.96	6.96	6.03	5.04	4.29	4.02
TIFP (PDB)	0.54	0.00	3.91	5.33	3.32	2.57	2.53	2.54
TM-align	0.66	6.52	6.96	6.96	5.76	4.64	3.89	3.57

Table A2.19. AUC and EFs of different binding site comparison methods for dataset 6.2.

Method	AUC	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	0.57	7.58	7.58	6.97	4.13	3.38	3.03	2.64
FuzCav	0.54	7.58	7.58	7.42	4.17	3.03	2.46	2.27
FuzCav (PDB)	0.54	7.58	7.58	7.42	4.17	3.08	2.50	2.27
Grim	0.61	7.58	7.58	4.39	3.60	2.80	2.84	2.71
Grim (PDB)	0.58	0.00	4.39	4.92	3.71	2.47	2.56	2.64
IsoMIF	0.81	7.58	7.58	7.58	7.42	6.89	6.16	5.97
KRIPO	0.77	7.58	7.58	7.58	5.91	5.35	4.62	4.27
PocketMatch	0.60	7.58	7.58	7.27	4.89	4.17	3.67	3.35
ProBiS	0.55	7.58	7.42	7.50	5.04	3.66	3.26	2.92
RAPMAD	0.52	7.58	7.58	7.27	4.36	3.08	2.44	2.14
Shaper	0.65	7.58	7.58	7.35	5.15	4.29	3.71	3.35
Shaper (PDB)	0.64	7.58	7.58	7.35	5.11	4.27	3.71	3.41
VolSite/ Shaper	0.58	7.58	7.58	6.36	4.24	3.33	2.92	2.64
VolSite/ Shaper (PDB)	0.56	7.58	7.58	5.15	3.33	2.83	2.58	2.41
SiteAlign	0.57	7.58	7.58	7.50	4.85	3.51	2.86	2.52
SiteEngine	0.57	7.58	7.58	6.97	4.89	3.81	3.35	3.02
SiteHopper	0.81	7.58	7.58	7.58	6.97	6.62	6.36	5.80
SMAP	0.54	7.58	7.58	7.42	4.73	3.36	2.90	2.58
TIFP	0.63	7.58	7.58	6.97	5.49	4.62	3.96	3.71
TIFP (PDB)	0.53	0.00	5.45	4.17	2.77	2.73	2.42	2.33
TM-align	0.62	7.58	7.58	7.58	5.34	4.19	3.71	3.26

Table A2.20. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 6. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.58 - 0.63	0.54 - 0.58	0.53 - 0.58	0.59 - 0.63	0.67 - 0.71	0.73 - 0.78	0.75 - 0.78	0.64 - 0.68	0.53 - 0.56	0.53 - 0.57
Cavbase	0.00	-0.05	-0.05	0.00	0.08	0.15	0.16	0.06	-0.06	-0.05
	1.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00
FuzCav (PDB)	0.05	0.00	0.00	0.05	0.13	0.20	0.21	0.11	-0.01	-0.01
	0.00	1.00	0.94	0.00	0.00	0.00	0.00	0.00	0.33	0.67
FuzCav	0.05	0.00	0.00	0.05	0.13	0.20	0.21	0.11	-0.01	-0.01
	0.00	0.94	1.00	0.00	0.00	0.00	0.00	0.00	0.38	0.73
Grim (PDB)	0.00	-0.05	-0.05	0.00	0.08	0.15	0.16	0.06	-0.06	-0.06
	0.91	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Grim	-0.08	-0.13	-0.13	-0.08	0.00	0.07	0.08	-0.03	-0.15	-0.14
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.09	0.00	0.00
IsoMIF	-0.15	-0.20	-0.20	-0.15	-0.07	0.00	0.01	-0.09	-0.21	-0.20
	0.00	0.00	0.00	0.00	0.00	1.00	0.49	0.00	0.00	0.00
KRIPO	-0.16	-0.21	-0.21	-0.16	-0.08	-0.01	0.00	-0.10	-0.22	-0.21
	0.00	0.00	0.00	0.00	0.00	0.49	1.00	0.00	0.00	0.00
Pocket-Match	-0.06	-0.11	-0.11	-0.06	0.03	0.09	0.10	0.00	-0.12	-0.11
	0.00	0.00	0.00	0.00	0.09	0.00	0.00	1.00	0.00	0.00
ProBiS	0.06	0.01	0.01	0.06	0.15	0.21	0.22	0.12	0.00	0.01
	0.00	0.33	0.38	0.00	0.00	0.00	0.00	0.00	1.00	0.63
RAPMAD	0.05	0.01	0.01	0.06	0.14	0.20	0.21	0.11	-0.01	0.00
	0.00	0.67	0.73	0.00	0.00	0.00	0.00	0.00	0.63	1.00
VolSite/ Shaper (PDB)	0.04	-0.01	-0.01	0.04	0.12	0.19	0.20	0.09	-0.03	-0.02
	0.02	0.40	0.36	0.01	0.00	0.00	0.00	0.00	0.05	0.21
VolSite/ Shaper	0.04	0.00	-0.01	0.05	0.13	0.19	0.20	0.10	-0.02	-0.01
	0.00	0.77	0.71	0.00	0.00	0.00	0.00	0.00	0.19	0.48
Shaper (PDB)	-0.06	-0.11	-0.11	-0.06	0.02	0.09	0.10	0.00	-0.12	-0.11
	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.93	0.00	0.00
Shaper	-0.05	-0.10	-0.10	-0.05	0.03	0.10	0.11	0.01	-0.11	-0.10
	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.52	0.00	0.00
SiteAlign	0.02	-0.03	-0.03	0.02	0.10	0.17	0.18	0.08	-0.04	-0.04
	0.26	0.04	0.03	0.20	0.00	0.00	0.00	0.00	0.00	0.01
SiteEngine	-0.03	-0.08	-0.08	-0.03	0.05	0.12	0.13	0.03	-0.09	-0.09
	0.04	0.00	0.00	0.05	0.00	0.00	0.00	0.07	0.00	0.00
SiteHopper	-0.12	-0.17	-0.17	-0.12	-0.04	0.03	0.04	-0.06	-0.18	-0.17
	0.00	0.00	0.00	0.00	0.01	0.03	0.00	0.00	0.00	0.00
SMAP	-0.01	-0.06	-0.06	-0.01	0.07	0.14	0.15	0.04	-0.08	-0.07
	0.34	0.00	0.00	0.39	0.00	0.00	0.00	0.00	0.00	0.00
TIFP (PDB)	0.07	0.02	0.02	0.07	0.15	0.22	0.23	0.12	0.00	0.01
	0.00	0.21	0.24	0.00	0.00	0.00	0.00	0.00	0.71	0.43
TIFP	-0.10	-0.15	-0.15	-0.10	-0.02	0.05	0.06	-0.04	-0.16	-0.16
	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00
TM-align	-0.05	-0.10	-0.10	-0.05	0.03	0.10	0.10	0.00	-0.12	-0.11
	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.80	0.00	0.00

Table A2.20 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 6. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.55 - 0.59	0.54 - 0.58	0.64 - 0.69	0.63 - 0.67	0.57 - 0.61	0.62 - 0.66	0.71 - 0.74	0.60 - 0.64	0.52 - 0.56	0.69 - 0.73	0.64 - 0.68
Cavbase	-0.04	-0.04	0.06	0.05	-0.02	0.03	0.12	0.01	-0.07	0.10	0.05
	0.02	0.00	0.00	0.00	0.26	0.04	0.00	0.34	0.00	0.00	0.00
FuzCav (PDB)	0.01	0.00	0.11	0.10	0.03	0.08	0.17	0.06	-0.02	0.15	0.10
	0.40	0.77	0.00	0.00	0.04	0.00	0.00	0.00	0.21	0.00	0.00
FuzCav	0.01	0.01	0.11	0.10	0.03	0.08	0.17	0.06	-0.02	0.15	0.10
	0.36	0.71	0.00	0.00	0.03	0.00	0.00	0.00	0.24	0.00	0.00
Grim (PDB)	-0.04	-0.05	0.06	0.05	-0.02	0.03	0.12	0.01	-0.07	0.10	0.05
	0.01	0.00	0.00	0.00	0.20	0.05	0.00	0.39	0.00	0.00	0.00
Grim	-0.12	-0.13	-0.02	-0.03	-0.10	-0.05	0.04	-0.07	-0.15	0.02	-0.03
	0.00	0.00	0.12	0.02	0.00	0.00	0.01	0.00	0.00	0.19	0.05
IsoMIF	-0.19	-0.19	-0.09	-0.10	-0.17	-0.12	-0.03	-0.14	-0.22	-0.05	-0.10
	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
KRIPO	-0.20	-0.20	-0.10	-0.11	-0.18	-0.13	-0.04	-0.15	-0.23	-0.06	-0.10
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PocketMatch	-0.09	-0.10	0.00	-0.01	-0.08	-0.03	0.06	-0.04	-0.12	0.04	0.00
	0.00	0.00	0.93	0.52	0.00	0.07	0.00	0.00	0.00	0.00	0.80
ProBiS	0.03	0.02	0.12	0.11	0.04	0.09	0.18	0.08	0.00	0.16	0.12
	0.05	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.00	0.00
RAPMAD	0.02	0.01	0.11	0.10	0.04	0.09	0.17	0.07	-0.01	0.16	0.11
	0.21	0.48	0.00	0.00	0.01	0.00	0.00	0.00	0.43	0.00	0.00
VolSite/ Shaper (PDB)	0.00	-0.01	0.10	0.08	0.02	0.07	0.16	0.05	-0.03	0.14	0.09
	1.00	0.58	0.00	0.00	0.21	0.00	0.00	0.00	0.03	0.00	0.00
VolSite/ Shaper	0.01	0.00	0.10	0.09	0.03	0.08	0.16	0.06	-0.02	0.15	0.10
	0.58	1.00	0.00	0.00	0.07	0.00	0.00	0.00	0.11	0.00	0.00
Shaper (PDB)	-0.10	-0.10	0.00	-0.01	-0.08	-0.03	0.06	-0.05	-0.13	0.04	-0.01
	0.00	0.00	1.00	0.48	0.00	0.06	0.00	0.00	0.00	0.00	0.74
Shaper	-0.08	-0.09	0.01	0.00	-0.07	-0.02	0.07	-0.03	-0.12	0.05	0.01
	0.00	0.00	0.48	1.00	0.00	0.25	0.00	0.02	0.00	0.00	0.70
SiteAlign	-0.02	-0.03	0.08	0.07	0.00	0.05	0.14	0.03	-0.05	0.12	0.07
	0.21	0.07	0.00	0.00	1.00	0.00	0.00	0.03	0.00	0.00	0.00
SiteEngine	-0.07	-0.08	0.03	0.02	-0.05	0.00	0.09	-0.02	-0.10	0.07	0.02
	0.00	0.00	0.06	0.25	0.00	1.00	0.00	0.24	0.00	0.00	0.11
SiteHopper	-0.16	-0.16	-0.06	-0.07	-0.14	-0.09	0.00	-0.11	-0.19	-0.02	-0.06
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.23	0.00
SMAP	-0.05	-0.06	0.05	0.03	-0.03	0.02	0.11	0.00	-0.08	0.09	0.04
	0.00	0.00	0.00	0.02	0.03	0.24	0.00	1.00	0.00	0.00	0.01
TIFP (PDB)	0.03	0.02	0.13	0.12	0.05	0.10	0.19	0.08	0.00	0.17	0.12
	0.03	0.11	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
TIFP	-0.14	-0.15	-0.04	-0.05	-0.12	-0.07	0.02	-0.09	-0.17	0.00	-0.05
	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	1.00	0.00
TM-align	-0.09	-0.10	0.01	-0.01	-0.07	-0.02	0.06	-0.04	-0.12	0.05	0.00
	0.00	0.00	0.74	0.70	0.00	0.11	0.00	0.01	0.00	0.00	1.00

Table A2.21. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 6.2. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.55 - 0.58	0.52 - 0.56	0.52 - 0.55	0.56 - 0.59	0.60 - 0.63	0.79 - 0.83	0.76 - 0.79	0.58 - 0.62	0.53 - 0.56	0.50 - 0.53
Cavbase	0.00	-0.03	-0.03	0.01	0.05	0.25	0.21	0.03	-0.02	-0.05
	1.00	0.04	0.03	0.34	0.00	0.00	0.00	0.01	0.11	0.00
FuzCav (PDB)	0.03	0.00	0.00	0.04	0.08	0.27	0.24	0.06	0.01	-0.02
	0.04	1.00	0.95	0.00	0.00	0.00	0.00	0.00	0.46	0.09
FuzCav	0.03	0.00	0.00	0.04	0.08	0.27	0.24	0.06	0.01	-0.02
	0.03	0.95	1.00	0.00	0.00	0.00	0.00	0.00	0.42	0.10
Grim (PDB)	-0.01	-0.04	-0.04	0.00	0.04	0.23	0.20	0.02	-0.03	-0.06
	0.34	0.00	0.00	1.00	0.00	0.00	0.00	0.08	0.00	0.00
Grim	-0.05	-0.08	-0.08	-0.04	0.00	0.20	0.16	-0.01	-0.07	-0.10
	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.26	0.00	0.00
IsoMIF	-0.25	-0.27	-0.27	-0.23	-0.20	0.00	-0.04	-0.21	-0.26	-0.30
	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
KRIPO	-0.21	-0.24	-0.24	-0.20	-0.16	0.04	0.00	-0.18	-0.23	-0.26
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
PocketMatch	-0.03	-0.06	-0.06	-0.02	0.01	0.21	0.18	0.00	-0.05	-0.08
	0.01	0.00	0.00	0.08	0.26	0.00	0.00	1.00	0.00	0.00
ProBiS	0.02	-0.01	-0.01	0.03	0.07	0.26	0.23	0.05	0.00	-0.03
	0.11	0.46	0.42	0.00	0.00	0.00	0.00	0.00	1.00	0.01
RAPMAD	0.05	0.02	0.02	0.06	0.10	0.30	0.26	0.08	0.03	0.00
	0.00	0.09	0.10	0.00	0.00	0.00	0.00	0.00	0.01	1.00
VolSite/ Shaper (PDB)	0.01	-0.02	-0.02	0.02	0.06	0.26	0.22	0.04	-0.01	-0.04
	0.44	0.13	0.12	0.06	0.00	0.00	0.00	0.00	0.36	0.00
VolSite/ Shaper	-0.02	-0.04	-0.04	0.00	0.03	0.23	0.19	0.02	-0.03	-0.06
	0.25	0.00	0.00	0.79	0.01	0.00	0.00	0.15	0.00	0.00
Shaper (PDB)	-0.08	-0.11	-0.11	-0.07	-0.03	0.17	0.13	-0.04	-0.10	-0.13
	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
Shaper	-0.08	-0.11	-0.11	-0.07	-0.04	0.16	0.13	-0.05	-0.10	-0.13
	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
SiteAlign	0.00	-0.03	-0.03	0.01	0.05	0.24	0.21	0.03	-0.02	-0.05
	0.81	0.02	0.01	0.47	0.00	0.00	0.00	0.02	0.05	0.00
SiteEngine	-0.01	-0.03	-0.03	0.01	0.04	0.24	0.20	0.03	-0.03	-0.06
	0.65	0.01	0.01	0.65	0.00	0.00	0.00	0.04	0.04	0.00
SiteHopper	-0.25	-0.27	-0.28	-0.24	-0.20	0.00	-0.04	-0.21	-0.27	-0.30
	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00
SMAP	0.03	0.00	0.00	0.04	0.07	0.27	0.24	0.06	0.01	-0.02
	0.05	0.94	0.90	0.00	0.00	0.00	0.00	0.00	0.52	0.08
TIFP (PDB)	0.03	0.01	0.01	0.05	0.08	0.28	0.24	0.07	0.02	-0.01
	0.00	0.51	0.55	0.00	0.00	0.00	0.00	0.00	0.10	0.19
TIFP	-0.06	-0.09	-0.09	-0.05	-0.01	0.18	0.15	-0.03	-0.08	-0.11
	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.03	0.00	0.00
TM-align	-0.06	-0.09	-0.09	-0.05	-0.01	0.19	0.15	-0.02	-0.08	-0.11
	0.00	0.00	0.00	0.00	0.44	0.00	0.00	0.06	0.00	0.00

Table 2.21 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 6.2. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.54 - 0.57	0.56 - 0.60	0.63 - 0.66	0.63 - 0.67	0.55 - 0.59	0.55 - 0.59	0.80 - 0.82	0.52 - 0.56	0.52 - 0.54	0.61 - 0.65	0.61 - 0.64
Cavbase	-0.01	0.02	0.08	0.08	0.00	0.01	0.25	-0.03	-0.03	0.06	0.06
	0.44	0.25	0.00	0.00	0.81	0.65	0.00	0.05	0.00	0.00	0.00
FuzCav (PDB)	0.02	0.04	0.11	0.11	0.03	0.03	0.27	0.00	-0.01	0.09	0.09
	0.13	0.00	0.00	0.00	0.02	0.01	0.00	0.94	0.51	0.00	0.00
FuzCav	0.02	0.04	0.11	0.11	0.03	0.03	0.28	0.00	-0.01	0.09	0.09
	0.12	0.00	0.00	0.00	0.01	0.01	0.00	0.90	0.55	0.00	0.00
Grim (PDB)	-0.02	0.00	0.07	0.07	-0.01	-0.01	0.24	-0.04	-0.05	0.05	0.05
	0.06	0.79	0.00	0.00	0.47	0.65	0.00	0.00	0.00	0.00	0.00
Grim	-0.06	-0.03	0.03	0.04	-0.05	-0.04	0.20	-0.07	-0.08	0.01	0.01
	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.30	0.44
IsoMIF	-0.26	-0.23	-0.17	-0.16	-0.24	-0.24	0.00	-0.27	-0.28	-0.18	-0.19
	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00
KRIPO	-0.22	-0.19	-0.13	-0.13	-0.21	-0.20	0.04	-0.24	-0.24	-0.15	-0.15
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PocketMatch	-0.04	-0.02	0.04	0.05	-0.03	-0.03	0.21	-0.06	-0.07	0.03	0.02
	0.00	0.15	0.00	0.00	0.02	0.04	0.00	0.00	0.00	0.03	0.06
ProBiS	0.01	0.03	0.10	0.10	0.02	0.03	0.27	-0.01	-0.02	0.08	0.08
	0.36	0.00	0.00	0.00	0.05	0.04	0.00	0.52	0.10	0.00	0.00
RAPMAD	0.04	0.06	0.13	0.13	0.05	0.06	0.30	0.02	0.01	0.11	0.11
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.19	0.00	0.00
VolSite/ Shaper (PDB)	0.00	0.02	0.09	0.09	0.01	0.02	0.26	-0.02	-0.03	0.07	0.07
	1.00	0.04	0.00	0.00	0.29	0.21	0.00	0.17	0.01	0.00	0.00
VolSite/ Shaper	-0.02	0.00	0.06	0.07	-0.01	-0.01	0.23	-0.04	-0.05	0.05	0.04
	0.04	1.00	0.00	0.00	0.34	0.50	0.00	0.00	0.00	0.00	0.00
Shaper (PDB)	-0.09	-0.06	0.00	0.01	-0.07	-0.07	0.17	-0.10	-0.11	-0.02	-0.02
	0.00	0.00	1.00	0.67	0.00	0.00	0.00	0.00	0.00	0.22	0.13
Shaper	-0.09	-0.07	-0.01	0.00	-0.08	-0.08	0.16	-0.11	-0.12	-0.02	-0.03
	0.00	0.00	0.67	1.00	0.00	0.00	0.00	0.00	0.00	0.09	0.05
SiteAlign	-0.01	0.01	0.07	0.08	0.00	0.00	0.24	-0.03	-0.04	0.06	0.06
	0.29	0.34	0.00	0.00	1.00	0.82	0.00	0.02	0.00	0.00	0.00
SiteEngine	-0.02	0.01	0.07	0.08	0.00	0.00	0.24	-0.03	-0.04	0.06	0.05
	0.21	0.50	0.00	0.00	0.82	1.00	0.00	0.02	0.00	0.00	0.00
SiteHopper	-0.26	-0.23	-0.17	-0.16	-0.24	-0.24	0.00	-0.27	-0.28	-0.19	-0.19
	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
SMAP	0.02	0.04	0.10	0.11	0.03	0.03	0.27	0.00	-0.01	0.09	0.08
	0.17	0.00	0.00	0.00	0.02	0.02	0.00	1.00	0.47	0.00	0.00
TIFP (PDB)	0.03	0.05	0.11	0.12	0.04	0.04	0.28	0.01	0.00	0.10	0.09
	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.47	1.00	0.00	0.00
TIFP	-0.07	-0.05	0.02	0.02	-0.06	-0.06	0.19	-0.09	-0.10	0.00	0.00
	0.00	0.00	0.22	0.09	0.00	0.00	0.00	0.00	0.00	1.00	0.78
TM-align	-0.07	-0.04	0.02	0.03	-0.06	-0.05	0.19	-0.08	-0.09	0.00	0.00
	0.00	0.00	0.13	0.05	0.00	0.00	0.00	0.00	0.00	0.78	1.00

Table A2.22. AUC and EFs of different binding site comparison methods for dataset 7.

Method	AUC	$EF_{1.6\%}$	$EF_{8.1\%}$	$EF_{16.1\%}$	$EF_{32.3\%}$	$EF_{48.4\%}$	$EF_{64.5\%}$	$EF_{80.6\%}$
Cavbase	0.55	3.26	0.65	1.31	0.82	1.09	0.98	1.04
FuzCav	0.67	3.26	1.31	0.98	1.31	1.31	1.22	1.24
FuzCav (PDB)	0.65	3.26	1.31	0.98	1.31	1.31	1.14	1.24
Grim	0.45	0.00	0.00	0.33	0.65	0.87	0.98	0.85
Grim (PDB)	0.45	0.00	0.00	0.33	0.82	0.98	0.82	0.65
IsoMIF	0.62	0.00	0.65	1.31	1.31	1.09	1.06	1.24
KRIPO	0.73	0.00	0.00	0.00	2.63	1.75	1.32	2.11
PocketMatch	0.51	0.00	0.65	0.98	0.98	0.98	0.73	0.59
ProBiS	0.50	3.26	0.65	0.33	0.16	0.11	0.08	0.52
RAPMAD	0.60	0.00	0.00	0.33	1.14	1.09	1.14	1.17
Shaper	0.54	3.26	1.31	0.65	0.65	0.98	1.06	1.17
Shaper (PDB)	0.54	3.26	1.31	0.65	0.65	0.98	1.14	1.17
VolSite/ Shaper	0.71	3.26	1.31	1.31	1.47	1.41	1.22	1.24
VolSite/ Shaper (PDB)	0.50	0.00	0.65	0.98	0.98	0.98	0.90	0.72
SiteAlign	0.44	0.00	0.00	0.33	0.65	0.76	0.98	1.04
SiteEngine	0.55	0.00	0.65	0.65	0.98	1.09	1.06	1.11
SiteHopper	0.56	0.00	0.65	0.65	1.14	1.09	1.14	1.11
SMAP	0.68	3.26	2.61	2.28	1.63	1.31	1.14	0.98
TIFP	0.55	3.26	0.65	0.33	0.82	1.20	1.06	0.85
TIFP (PDB)	0.56	3.26	0.65	0.98	1.31	0.98	0.82	0.65
TM-align	0.59	3.26	0.65	1.31	1.14	0.98	1.06	1.17

Table A2.23. AUC and EFs of different binding site comparison methods for dataset 7.2.

Method	AUC	$EF_{1.6\%}$	$EF_{8.1\%}$	$EF_{16.1\%}$	$EF_{32.3\%}$	$EF_{48.4\%}$	$EF_{64.5\%}$	$EF_{80.6\%}$
Cavbase	0.55	3.26	0.65	1.31	0.82	1.09	1.06	1.04
FuzCav	0.73	3.26	1.31	0.98	1.47	1.41	1.39	1.24
FuzCav (PDB)	0.72	3.26	1.31	0.98	1.47	1.41	1.39	1.24
Grim	0.65	0.00	0.65	0.98	1.14	1.20	1.39	1.24
Grim (PDB)	0.50	0.00	0.00	0.65	0.82	1.09	0.82	0.65
IsoMIF	0.62	0.00	0.65	1.31	1.31	1.09	1.06	1.24
KRIPO	0.74	0.00	0.00	0.00	2.63	1.75	1.32	2.11
PocketMatch	0.51	0.00	0.65	0.98	0.98	0.98	0.73	0.59
ProBiS	0.50	3.26	0.65	0.33	0.16	0.11	0.08	0.52
RAPMAD	0.60	0.00	0.00	0.65	1.14	1.09	1.14	1.17
Shaper	0.65	3.26	1.31	0.65	0.98	1.20	1.31	1.24
Shaper (PDB)	0.65	3.26	1.31	0.65	0.98	1.20	1.31	1.24
VolSite/ Shaper	0.76	3.26	1.31	1.96	1.47	1.63	1.39	1.24
VolSite/ Shaper (PDB)	0.57	0.00	0.65	1.31	1.14	1.09	0.90	0.72
SiteAlign	0.56	0.00	0.65	0.33	0.98	0.98	1.31	1.24
SiteEngine	0.55	0.00	0.65	0.65	0.98	1.09	1.06	1.11
SiteHopper	0.54	0.00	0.65	0.65	1.14	0.98	1.06	1.11
SMAP	0.68	3.26	2.61	2.28	1.63	1.31	1.14	0.98
TIFP	0.60	3.26	0.65	0.33	1.14	1.31	1.06	0.85
TIFP (PDB)	0.61	3.26	1.31	1.63	1.31	0.87	0.65	0.65
TM-align	0.59	3.26	0.65	1.31	1.14	0.98	1.06	1.17

Table A2.24. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 7. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.39 - 0.70	0.52 - 0.79	0.53 - 0.80	0.30 - 0.59	0.31 - 0.59	0.47 - 0.77	0.59 - 0.87	0.36 - 0.66	0.40 - 0.59	0.46 - 0.74
Cavbase	0.00	0.11	0.12	-0.10	-0.10	0.07	0.19	-0.03	-0.05	0.05
	1.00	0.31	0.25	0.35	0.36	0.51	0.08	0.76	0.58	0.63
FuzCav (PDB)	-0.11	0.00	0.01	-0.21	-0.20	-0.04	0.08	-0.14	-0.16	-0.06
	0.31	1.00	0.89	0.04	0.05	0.73	0.44	0.18	0.07	0.58
FuzCav	-0.12	-0.01	0.00	-0.22	-0.22	-0.05	0.06	-0.15	-0.17	-0.07
	0.25	0.89	1.00	0.03	0.03	0.63	0.52	0.14	0.05	0.49
Grim (PDB)	0.10	0.21	0.22	0.00	0.00	0.17	0.29	0.07	0.05	0.15
	0.35	0.04	0.03	1.00	0.98	0.11	0.01	0.53	0.58	0.15
Grim	0.10	0.20	0.22	0.00	0.00	0.17	0.28	0.06	0.05	0.15
	0.36	0.05	0.03	0.98	1.00	0.11	0.01	0.55	0.60	0.15
IsoMIF	-0.07	0.04	0.05	-0.17	-0.17	0.00	0.11	-0.10	-0.12	-0.02
	0.51	0.73	0.63	0.11	0.11	1.00	0.28	0.33	0.18	0.85
KRIPO	-0.19	-0.08	-0.06	-0.29	-0.28	-0.11	0.00	-0.22	-0.24	-0.13
	0.08	0.44	0.52	0.01	0.01	0.28	1.00	0.04	0.01	0.19
PocketMatch	0.03	0.14	0.15	-0.07	-0.06	0.10	0.22	0.00	-0.02	0.08
	0.76	0.18	0.14	0.53	0.55	0.33	0.04	1.00	0.85	0.42
ProBiS	0.05	0.16	0.17	-0.05	-0.05	0.12	0.24	0.02	0.00	0.10
	0.58	0.07	0.05	0.58	0.60	0.18	0.01	0.85	1.00	0.24
RAPMAD	-0.05	0.06	0.07	-0.15	-0.15	0.02	0.13	-0.08	-0.10	0.00
	0.63	0.58	0.49	0.15	0.15	0.85	0.19	0.42	0.24	1.00
VolSite/ Shaper (PDB)	0.05	0.16	0.17	-0.05	-0.05	0.12	0.24	0.02	0.00	0.10
	0.65	0.14	0.11	0.64	0.66	0.27	0.03	0.88	0.99	0.34
VolSite/ Shaper	-0.17	-0.06	-0.04	-0.27	-0.26	-0.09	0.02	-0.20	-0.22	-0.11
	0.11	0.55	0.64	0.01	0.01	0.35	0.84	0.05	0.01	0.25
Shaper (PDB)	0.00	0.11	0.12	-0.10	-0.09	0.07	0.19	-0.03	-0.05	0.05
	0.98	0.29	0.23	0.36	0.37	0.49	0.07	0.78	0.59	0.61
Shaper	0.01	0.11	0.13	-0.09	-0.09	0.08	0.19	-0.03	-0.05	0.06
	0.96	0.28	0.22	0.37	0.38	0.48	0.07	0.79	0.61	0.59
SiteAlign	0.11	0.22	0.23	0.01	0.01	0.18	0.30	0.08	0.06	0.16
	0.31	0.04	0.03	0.93	0.90	0.09	0.01	0.48	0.51	0.13
SiteEngine	0.00	0.10	0.12	-0.11	-0.10	0.07	0.18	-0.04	-0.06	0.05
	0.96	0.33	0.27	0.33	0.34	0.54	0.09	0.72	0.54	0.66
SiteHopper	-0.02	0.09	0.10	-0.12	-0.11	0.05	0.17	-0.05	-0.07	0.03
	0.88	0.40	0.33	0.28	0.29	0.62	0.12	0.65	0.46	0.75
SMAP	-0.13	-0.02	-0.01	-0.23	-0.23	-0.06	0.06	-0.16	-0.18	-0.08
	0.26	0.83	0.93	0.05	0.05	0.61	0.63	0.16	0.07	0.48
TIFP (PDB)	-0.01	0.10	0.11	-0.11	-0.11	0.06	0.17	-0.04	-0.06	0.04
	0.92	0.36	0.30	0.30	0.31	0.58	0.10	0.68	0.50	0.71
TIFP	0.00	0.11	0.12	-0.10	-0.10	0.07	0.18	-0.03	-0.05	0.05
	1.00	0.31	0.25	0.35	0.36	0.51	0.08	0.75	0.57	0.63
TM-align	-0.04	0.06	0.08	-0.14	-0.14	0.03	0.14	-0.08	-0.09	0.01
	0.70	0.54	0.46	0.19	0.19	0.79	0.18	0.49	0.31	0.94

Table A2.24 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 7. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.34 - 0.65	0.58 - 0.84	0.40 - 0.69	0.39 - 0.69	0.29 - 0.58	0.40 - 0.70	0.41 - 0.72	0.51 - 0.85	0.40 - 0.71	0.40 - 0.70	0.44 - 0.74
Cavbase	-0.05	0.17	0.00	-0.01	-0.11	0.00	0.02	0.13	0.01	0.00	0.04
	0.65	0.11	0.98	0.96	0.31	0.96	0.88	0.26	0.92	1.00	0.70
FuzCav (PDB)	-0.16	0.06	-0.11	-0.11	-0.22	-0.10	-0.09	0.02	-0.10	-0.11	-0.06
	0.14	0.55	0.29	0.28	0.04	0.33	0.40	0.83	0.36	0.31	0.54
FuzCav	-0.17	0.04	-0.12	-0.13	-0.23	-0.12	-0.10	0.01	-0.11	-0.12	-0.08
	0.11	0.64	0.23	0.22	0.03	0.27	0.33	0.93	0.30	0.25	0.46
Grim (PDB)	0.05	0.27	0.10	0.09	-0.01	0.11	0.12	0.23	0.11	0.10	0.14
	0.64	0.01	0.36	0.37	0.93	0.33	0.28	0.05	0.30	0.35	0.19
Grim	0.05	0.26	0.09	0.09	-0.01	0.10	0.11	0.23	0.11	0.10	0.14
	0.66	0.01	0.37	0.38	0.90	0.34	0.29	0.05	0.31	0.36	0.19
IsoMIF	-0.12	0.09	-0.07	-0.08	-0.18	-0.07	-0.05	0.06	-0.06	-0.07	-0.03
	0.27	0.35	0.49	0.48	0.09	0.54	0.62	0.61	0.58	0.51	0.79
KRIPO	-0.24	-0.02	-0.19	-0.19	-0.30	-0.18	-0.17	-0.06	-0.17	-0.18	-0.14
	0.03	0.84	0.07	0.07	0.01	0.09	0.12	0.63	0.10	0.08	0.18
PocketMatch	-0.02	0.20	0.03	0.03	-0.08	0.04	0.05	0.16	0.04	0.03	0.08
	0.88	0.05	0.78	0.79	0.48	0.72	0.65	0.16	0.68	0.75	0.49
ProBiS	0.00	0.22	0.05	0.05	-0.06	0.06	0.07	0.18	0.06	0.05	0.09
	0.99	0.01	0.59	0.61	0.51	0.54	0.46	0.07	0.50	0.57	0.31
RAPMAD	-0.10	0.11	-0.05	-0.06	-0.16	-0.05	-0.03	0.08	-0.04	-0.05	-0.01
	0.34	0.25	0.61	0.59	0.13	0.66	0.75	0.48	0.71	0.63	0.94
VolSite/ Shaper (PDB)	0.00	0.21	0.05	0.04	-0.06	0.05	0.07	0.18	0.06	0.05	0.09
	1.00	0.04	0.67	0.68	0.58	0.62	0.55	0.12	0.58	0.65	0.41
VolSite/ Shaper	-0.21	0.00	-0.17	-0.17	-0.28	-0.16	-0.15	-0.03	-0.15	-0.16	-0.12
	0.04	1.00	0.10	0.09	0.01	0.12	0.15	0.75	0.14	0.11	0.23
Shaper (PDB)	-0.05	0.17	0.00	0.00	-0.11	0.01	0.02	0.13	0.01	0.00	0.05
	0.67	0.10	1.00	0.98	0.32	0.94	0.85	0.25	0.90	0.97	0.68
Shaper	-0.04	0.17	0.00	0.00	-0.10	0.01	0.02	0.14	0.02	0.01	0.05
	0.68	0.09	0.98	1.00	0.33	0.92	0.84	0.24	0.88	0.95	0.66
SiteAlign	0.06	0.28	0.11	0.10	0.00	0.12	0.13	0.24	0.12	0.11	0.15
	0.58	0.01	0.32	0.33	1.00	0.29	0.25	0.04	0.27	0.30	0.16
SiteEngine	-0.05	0.16	-0.01	-0.01	-0.12	0.00	0.01	0.13	0.01	0.00	0.04
	0.62	0.12	0.94	0.92	0.29	1.00	0.91	0.28	0.96	0.97	0.73
SiteHopper	-0.07	0.15	-0.02	-0.02	-0.13	-0.01	0.00	0.11	-0.01	-0.02	0.03
	0.55	0.15	0.85	0.84	0.25	0.91	1.00	0.33	0.96	0.88	0.82
SMAP	-0.18	0.03	-0.13	-0.14	-0.24	-0.13	-0.11	0.00	-0.12	-0.13	-0.09
	0.12	0.75	0.25	0.24	0.04	0.28	0.33	1.00	0.31	0.26	0.45
TIFP (PDB)	-0.06	0.15	-0.01	-0.02	-0.12	-0.01	0.01	0.12	0.00	-0.01	0.03
	0.58	0.14	0.90	0.88	0.27	0.96	0.96	0.31	1.00	0.92	0.78
TIFP	-0.05	0.16	0.00	-0.01	-0.11	0.00	0.02	0.13	0.01	0.00	0.04
	0.65	0.11	0.97	0.95	0.30	0.97	0.88	0.26	0.92	1.00	0.70
TM-align	-0.09	0.12	-0.05	-0.05	-0.15	-0.04	-0.03	0.09	-0.03	-0.04	0.00
	0.41	0.23	0.68	0.66	0.16	0.73	0.82	0.45	0.78	0.70	1.00

Table A2.25. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 7.2. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.40 - 0.70	0.60 - 0.85	0.61 - 0.85	0.35 - 0.65	0.52 - 0.79	0.47 - 0.77	0.60 - 0.88	0.36 - 0.66	0.40 - 0.59	0.46 - 0.74
Cavbase	0.00	0.17	0.18	-0.05	0.10	0.07	0.19	-0.04	-0.05	0.05
	1.00	0.08	0.07	0.65	0.32	0.53	0.08	0.74	0.56	0.62
FuzCav (PDB)	-0.17	0.00	0.01	-0.22	-0.07	-0.11	0.01	-0.21	-0.23	-0.12
	0.08	1.00	0.92	0.02	0.45	0.29	0.89	0.04	0.01	0.21
FuzCav	-0.18	-0.01	0.00	-0.23	-0.08	-0.11	0.00	-0.22	-0.24	-0.13
	0.07	0.92	1.00	0.02	0.40	0.25	0.96	0.03	0.00	0.18
Grim (PDB)	0.05	0.22	0.23	0.00	0.15	0.12	0.24	0.01	0.00	0.10
	0.65	0.02	0.02	1.00	0.13	0.27	0.02	0.90	0.96	0.33
Grim	-0.10	0.07	0.08	-0.15	0.00	-0.04	0.08	-0.14	-0.16	-0.05
	0.32	0.45	0.40	0.13	1.00	0.73	0.40	0.18	0.06	0.61
IsoMIF	-0.07	0.11	0.11	-0.12	0.04	0.00	0.12	-0.10	-0.12	-0.02
	0.53	0.29	0.25	0.27	0.73	1.00	0.26	0.33	0.18	0.88
KRIPO	-0.19	-0.01	0.00	-0.24	-0.08	-0.12	0.00	-0.22	-0.24	-0.13
	0.08	0.89	0.96	0.02	0.40	0.26	1.00	0.04	0.01	0.19
PocketMatch	0.04	0.21	0.22	-0.01	0.14	0.10	0.22	0.00	-0.02	0.09
	0.74	0.04	0.03	0.90	0.18	0.33	0.04	1.00	0.85	0.40
ProBiS	0.05	0.23	0.24	0.00	0.16	0.12	0.24	0.02	0.00	0.11
	0.56	0.01	0.00	0.96	0.06	0.18	0.01	0.85	1.00	0.22
RAPMAD	-0.05	0.12	0.13	-0.10	0.05	0.02	0.13	-0.09	-0.11	0.00
	0.62	0.21	0.18	0.33	0.61	0.88	0.19	0.40	0.22	1.00
VolSite/ Shaper (PDB)	-0.02	0.16	0.17	-0.07	0.09	0.05	0.17	-0.05	-0.07	0.04
	0.87	0.11	0.09	0.53	0.40	0.64	0.11	0.62	0.43	0.73
VolSite/ Shaper	-0.22	-0.04	-0.03	-0.26	-0.11	-0.15	-0.03	-0.25	-0.27	-0.16
	0.03	0.64	0.71	0.01	0.23	0.13	0.77	0.01	0.00	0.09
Shaper (PDB)	-0.10	0.07	0.08	-0.15	0.00	-0.03	0.09	-0.14	-0.15	-0.05
	0.34	0.43	0.38	0.14	0.97	0.76	0.38	0.19	0.07	0.64
Shaper	-0.10	0.08	0.08	-0.15	0.00	-0.03	0.09	-0.14	-0.15	-0.05
	0.34	0.42	0.37	0.15	0.96	0.77	0.38	0.19	0.07	0.65
SiteAlign	-0.01	0.16	0.17	-0.06	0.09	0.05	0.17	-0.05	-0.07	0.04
	0.89	0.10	0.08	0.54	0.37	0.61	0.09	0.63	0.44	0.71
SiteEngine	0.00	0.17	0.18	-0.05	0.10	0.07	0.19	-0.04	-0.06	0.05
	0.98	0.09	0.07	0.63	0.33	0.54	0.08	0.72	0.54	0.63
SiteHopper	0.01	0.18	0.19	-0.04	0.11	0.08	0.20	-0.03	-0.05	0.06
	0.95	0.07	0.06	0.70	0.29	0.49	0.07	0.79	0.62	0.57
SMAP	-0.13	0.05	0.06	-0.18	-0.02	-0.06	0.06	-0.16	-0.18	-0.07
	0.28	0.67	0.61	0.12	0.83	0.61	0.60	0.16	0.07	0.51
TIFP (PDB)	-0.06	0.11	0.12	-0.11	0.04	0.01	0.12	-0.10	-0.12	-0.01
	0.56	0.25	0.21	0.28	0.68	0.96	0.23	0.35	0.19	0.92
TIFP	-0.05	0.12	0.13	-0.10	0.05	0.02	0.14	-0.09	-0.10	0.00
	0.64	0.20	0.17	0.35	0.59	0.86	0.18	0.42	0.25	0.97
TM-align	-0.04	0.13	0.14	-0.09	0.06	0.03	0.15	-0.08	-0.09	0.01
	0.72	0.18	0.16	0.41	0.54	0.79	0.17	0.49	0.31	0.90

Table A2.25 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 7.2. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.42 - 0.71	0.65 - 0.88	0.51 - 0.78	0.51 - 0.78	0.42 - 0.71	0.40 - 0.70	0.39 - 0.70	0.51 - 0.85	0.47 - 0.76	0.45 - 0.74	0.44 - 0.74
Cavbase	0.02	0.22	0.10	0.10	0.01	0.00	-0.01	0.13	0.06	0.05	0.04
	0.87	0.03	0.34	0.34	0.89	0.98	0.95	0.28	0.56	0.64	0.72
FuzCav (PDB)	-0.16	0.04	-0.07	-0.08	-0.16	-0.17	-0.18	-0.05	-0.11	-0.12	-0.13
	0.11	0.64	0.43	0.42	0.10	0.09	0.07	0.67	0.25	0.20	0.18
FuzCav	-0.17	0.03	-0.08	-0.08	-0.17	-0.18	-0.19	-0.06	-0.12	-0.13	-0.14
	0.09	0.71	0.38	0.37	0.08	0.07	0.06	0.61	0.21	0.17	0.16
Grim (PDB)	0.07	0.26	0.15	0.15	0.06	0.05	0.04	0.18	0.11	0.10	0.09
	0.53	0.01	0.14	0.15	0.54	0.63	0.70	0.12	0.28	0.35	0.41
Grim	-0.09	0.11	0.00	0.00	-0.09	-0.10	-0.11	0.02	-0.04	-0.05	-0.06
	0.40	0.23	0.97	0.96	0.37	0.33	0.29	0.83	0.68	0.59	0.54
IsoMIF	-0.05	0.15	0.03	0.03	-0.05	-0.07	-0.08	0.06	-0.01	-0.02	-0.03
	0.64	0.13	0.76	0.77	0.61	0.54	0.49	0.61	0.96	0.86	0.79
KRIPO	-0.17	0.03	-0.09	-0.09	-0.17	-0.19	-0.20	-0.06	-0.12	-0.14	-0.15
	0.11	0.77	0.38	0.38	0.09	0.08	0.07	0.60	0.23	0.18	0.17
PocketMatch	0.05	0.25	0.14	0.14	0.05	0.04	0.03	0.16	0.10	0.09	0.08
	0.62	0.01	0.19	0.19	0.63	0.72	0.79	0.16	0.35	0.42	0.49
ProBiS	0.07	0.27	0.15	0.15	0.07	0.06	0.05	0.18	0.12	0.10	0.09
	0.43	0.00	0.07	0.07	0.44	0.54	0.62	0.07	0.19	0.25	0.31
RAPMAD	-0.04	0.16	0.05	0.05	-0.04	-0.05	-0.06	0.07	0.01	0.00	-0.01
	0.73	0.09	0.64	0.65	0.71	0.63	0.57	0.51	0.92	0.97	0.90
VolSite/ Shaper (PDB)	0.00	0.20	0.08	0.08	0.00	-0.02	-0.03	0.11	0.05	0.03	0.02
	1.00	0.04	0.42	0.43	0.98	0.89	0.82	0.34	0.67	0.76	0.84
VolSite/ Shaper	-0.20	0.00	-0.12	-0.12	-0.20	-0.21	-0.22	-0.09	-0.15	-0.17	-0.18
	0.04	1.00	0.21	0.21	0.04	0.03	0.03	0.41	0.11	0.09	0.08
Shaper (PDB)	-0.08	0.12	0.00	0.00	-0.09	-0.10	-0.11	0.03	-0.04	-0.05	-0.06
	0.42	0.21	1.00	0.99	0.39	0.35	0.31	0.80	0.71	0.62	0.56
Shaper	-0.08	0.12	0.00	0.00	-0.08	-0.10	-0.11	0.03	-0.04	-0.05	-0.06
	0.43	0.21	0.99	1.00	0.40	0.35	0.31	0.80	0.72	0.62	0.57
SiteAlign	0.00	0.20	0.09	0.08	0.00	-0.01	-0.02	0.11	0.05	0.03	0.03
	0.98	0.04	0.39	0.40	1.00	0.91	0.84	0.32	0.64	0.74	0.81
SiteEngine	0.02	0.21	0.10	0.10	0.01	0.00	-0.01	0.13	0.06	0.05	0.04
	0.89	0.03	0.35	0.35	0.91	1.00	0.93	0.28	0.57	0.66	0.73
SiteHopper	0.03	0.22	0.11	0.11	0.02	0.01	0.00	0.14	0.07	0.06	0.05
	0.82	0.03	0.31	0.31	0.84	0.93	1.00	0.25	0.51	0.60	0.67
SMAP	-0.11	0.09	-0.03	-0.03	-0.11	-0.13	-0.14	0.00	-0.06	-0.08	-0.09
	0.34	0.41	0.80	0.80	0.32	0.28	0.25	1.00	0.57	0.49	0.45
TIFP (PDB)	-0.05	0.15	0.04	0.04	-0.05	-0.06	-0.07	0.06	0.00	-0.01	-0.02
	0.67	0.11	0.71	0.72	0.64	0.57	0.51	0.57	1.00	0.90	0.83
TIFP	-0.03	0.17	0.05	0.05	-0.03	-0.05	-0.06	0.08	0.01	0.00	-0.01
	0.76	0.09	0.62	0.62	0.74	0.66	0.60	0.49	0.90	1.00	0.93
TM-align	-0.02	0.18	0.06	0.06	-0.03	-0.04	-0.05	0.09	0.02	0.01	0.00
	0.84	0.08	0.56	0.57	0.81	0.73	0.67	0.45	0.83	0.93	1.00

Table A2.26. AUC and EFs of different binding site comparison methods for dataset 8.

Method	AUC	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	0.82	339.13	118.26	62.61	32.17	21.45	16.52	13.22
FuzCav	0.77	408.70	88.70	44.35	25.65	18.84	15.00	12.00
FuzCav (PDB)	0.77	408.70	85.22	44.35	24.78	18.26	15.00	12.00
Grim	0.70	313.04	76.52	40.00	20.87	14.20	11.09	9.39
Grim (PDB)	0.64	52.17	62.61	31.30	17.39	12.17	9.13	7.65
IsoMIF	0.87	408.70	107.83	60.00	30.87	21.16	15.87	13.04
KRIPO	0.85	426.09	106.09	56.52	30.87	20.58	16.30	13.04
PocketMatch	0.82	434.78	104.35	59.13	32.17	22.03	16.52	13.91
ProBiS	0.85	382.61	137.39	70.43	36.96	24.64	18.48	14.78
RAPMAD	0.74	400.00	83.48	41.74	20.87	13.91	10.43	8.70
Shaper	0.75	417.39	99.13	53.04	26.96	18.26	13.91	11.30
Shaper (PDB)	0.75	417.39	100.87	53.04	26.96	18.84	14.13	11.30
VolSite/ Shaper	0.77	408.70	107.83	55.65	27.83	18.84	14.35	11.65
VolSite/ Shaper (PDB)	0.72	400.00	102.61	52.17	26.09	17.39	13.04	10.61
SiteAlign	0.87	426.09	116.52	60.00	30.87	20.87	16.52	13.39
SiteEngine	0.86	373.91	118.26	65.22	32.61	22.03	16.74	13.57
SiteHopper	0.77	400.00	114.78	60.00	30.43	20.58	15.43	12.35
SMAP	0.86	313.04	118.26	65.22	34.35	23.19	17.83	14.43
TIFP	0.71	330.43	69.57	36.52	19.13	13.33	10.43	8.70
TIFP (PDB)	0.66	0.00	64.35	33.91	16.96	11.88	9.35	7.83
TM-align	0.88	434.78	135.65	68.70	37.83	26.38	19.78	15.83

Table A2.27. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 8. P-values below 0.05 are coloured green.

Method	Cavbase	FuzCav (PDB)	FuzCav	Grim (PDB)	Grim	IsoMIF	KRIPO	PocketMatch	ProBiS	RAPMAD
CI	0.76 - 0.87	0.70 - 0.83	0.71 - 0.83	0.57 - 0.71	0.63 - 0.77	0.83 - 0.92	0.80 - 0.89	0.76 - 0.88	0.80 - 0.90	0.68 - 0.80
Cavbase	0.00	-0.05	-0.05	-0.18	-0.12	0.05	0.03	0.00	0.03	-0.08
	1.00	0.25	0.27	0.00	0.01	0.15	0.47	0.95	0.42	0.05
FuzCav (PDB)	0.05	0.00	0.00	-0.13	-0.07	0.10	0.08	0.05	0.08	-0.03
	0.25	1.00	0.98	0.01	0.14	0.01	0.06	0.24	0.05	0.50
FuzCav	0.05	0.00	0.00	-0.13	-0.07	0.10	0.08	0.05	0.08	-0.03
	0.27	0.98	1.00	0.01	0.13	0.01	0.06	0.25	0.06	0.48
Grim (PDB)	0.18	0.13	0.13	0.00	0.06	0.23	0.21	0.18	0.21	0.10
	0.00	0.01	0.01	1.00	0.20	0.00	0.00	0.00	0.00	0.02
Grim	0.12	0.07	0.07	-0.06	0.00	0.17	0.15	0.12	0.15	0.04
	0.01	0.14	0.13	0.20	1.00	0.00	0.00	0.01	0.00	0.36
IsoMIF	-0.05	-0.10	-0.10	-0.23	-0.17	0.00	-0.03	-0.05	-0.02	-0.13
	0.15	0.01	0.01	0.00	0.00	1.00	0.44	0.18	0.52	0.00
KRIPO	-0.03	-0.08	-0.08	-0.21	-0.15	0.03	0.00	-0.02	0.00	-0.11
	0.47	0.06	0.06	0.00	0.00	0.44	1.00	0.53	0.91	0.00
PocketMatch	0.00	-0.05	-0.05	-0.18	-0.12	0.05	0.02	0.00	0.03	-0.08
	0.95	0.24	0.25	0.00	0.01	0.18	0.53	1.00	0.47	0.05
ProBiS	-0.03	-0.08	-0.08	-0.21	-0.15	0.02	0.00	-0.03	0.00	-0.11
	0.42	0.05	0.06	0.00	0.00	0.52	0.91	0.47	1.00	0.00
RAPMAD	0.08	0.03	0.03	-0.10	-0.04	0.13	0.11	0.08	0.11	0.00
	0.05	0.50	0.48	0.02	0.36	0.00	0.00	0.05	0.00	1.00
VolSite/ Shaper (PDB)	0.10	0.05	0.05	-0.08	-0.02	0.15	0.13	0.10	0.13	0.02
	0.02	0.28	0.27	0.10	0.71	0.00	0.00	0.02	0.00	0.62
VolSite/ Shaper	0.05	0.00	0.00	-0.13	-0.07	0.10	0.08	0.05	0.08	-0.03
	0.24	1.00	0.98	0.01	0.13	0.01	0.05	0.23	0.05	0.49
Shaper (PDB)	0.07	0.02	0.02	-0.11	-0.05	0.12	0.09	0.07	0.10	-0.01
	0.12	0.71	0.69	0.02	0.27	0.00	0.02	0.11	0.02	0.78
Shaper	0.06	0.01	0.02	-0.12	-0.06	0.12	0.09	0.07	0.09	-0.02
	0.14	0.76	0.74	0.01	0.24	0.00	0.03	0.13	0.02	0.72
SiteAlign	-0.05	-0.10	-0.10	-0.23	-0.17	0.00	-0.02	-0.05	-0.02	-0.13
	0.19	0.02	0.02	0.00	0.00	0.91	0.52	0.23	0.61	0.00
SiteEngine	-0.05	-0.10	-0.09	-0.23	-0.16	0.01	-0.02	-0.04	-0.02	-0.12
	0.21	0.02	0.02	0.00	0.00	0.84	0.57	0.25	0.66	0.00
SiteHopper	0.05	0.00	0.00	-0.14	-0.07	0.10	0.07	0.05	0.08	-0.03
	0.30	0.92	0.95	0.00	0.12	0.01	0.08	0.28	0.07	0.43
SMAP	-0.05	-0.10	-0.09	-0.23	-0.16	0.01	-0.02	-0.04	-0.02	-0.12
	0.22	0.02	0.02	0.00	0.00	0.84	0.58	0.26	0.67	0.00
TIFP (PDB)	0.16	0.11	0.11	-0.02	0.04	0.21	0.18	0.16	0.19	0.08
	0.00	0.02	0.02	0.60	0.42	0.00	0.00	0.00	0.00	0.07
TIFP	0.11	0.06	0.06	-0.07	-0.01	0.16	0.14	0.11	0.14	0.03
	0.01	0.18	0.18	0.15	0.87	0.00	0.00	0.01	0.00	0.46
TM-align	-0.06	-0.11	-0.11	-0.25	-0.18	-0.01	-0.04	-0.06	-0.03	-0.14
	0.08	0.00	0.01	0.00	0.00	0.71	0.27	0.10	0.33	0.00

Table A2.27 *(continued)*. AUC confidence intervals for the ROC curves of different binding site comparison methods and AUC value differences with the corresponding p-values calculated according to DeLong and co-workers[20] for dataset 8. P-values below 0.05 are coloured green.

Method	VolSite/ Shaper (PDB)	VolSite/ Shaper	Shaper (PDB)	Shaper	SiteAlign	SiteEngine	SiteHopper	SMAP	TIFP (PDB)	TIFP	TM-align
CI	0.65 - 0.79	0.71 - 0.83	0.69 - 0.82	0.69 - 0.82	0.82 - 0.91	0.82 - 0.91	0.71 - 0.84	0.82 - 0.91	0.60 - 0.72	0.64 - 0.77	0.84 - 0.93
Cavbase	-0.10	-0.05	-0.07	-0.06	0.05	0.05	-0.05	0.05	-0.16	-0.11	0.06
	0.02	0.24	0.12	0.14	0.19	0.21	0.30	0.22	0.00	0.01	0.08
FuzCav (PDB)	-0.05	0.00	-0.02	-0.01	0.10	0.10	0.00	0.10	-0.11	-0.06	0.11
	0.28	1.00	0.71	0.76	0.02	0.02	0.92	0.02	0.02	0.18	0.00
FuzCav	-0.05	0.00	-0.02	-0.02	0.10	0.09	0.00	0.09	-0.11	-0.06	0.11
	0.27	0.98	0.69	0.74	0.02	0.02	0.95	0.02	0.02	0.18	0.01
Grim (PDB)	0.08	0.13	0.11	0.12	0.23	0.23	0.14	0.23	0.02	0.07	0.25
	0.10	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.60	0.15	0.00
Grim	0.02	0.07	0.05	0.06	0.17	0.16	0.07	0.16	-0.04	0.01	0.18
	0.71	0.13	0.27	0.24	0.00	0.00	0.12	0.00	0.42	0.87	0.00
IsoMIF	-0.15	-0.10	-0.12	-0.12	0.00	-0.01	-0.10	-0.01	-0.21	-0.16	0.01
	0.00	0.01	0.00	0.00	0.91	0.84	0.01	0.84	0.00	0.00	0.71
KRIPO	-0.13	-0.08	-0.09	-0.09	0.02	0.02	-0.07	0.02	-0.18	-0.14	0.04
	0.00	0.05	0.02	0.03	0.52	0.57	0.08	0.58	0.00	0.00	0.27
PocketMatch	-0.10	-0.05	-0.07	-0.07	0.05	0.04	-0.05	0.04	-0.16	-0.11	0.06
	0.02	0.23	0.11	0.13	0.23	0.25	0.28	0.26	0.00	0.01	0.10
ProBiS	-0.13	-0.08	-0.10	-0.09	0.02	0.02	-0.08	0.02	-0.19	-0.14	0.03
	0.00	0.05	0.02	0.02	0.61	0.66	0.07	0.67	0.00	0.00	0.33
RAPMAD	-0.02	0.03	0.01	0.02	0.13	0.12	0.03	0.12	-0.08	-0.03	0.14
	0.62	0.49	0.78	0.72	0.00	0.00	0.43	0.00	0.07	0.46	0.00
VolSite/ Shaper (PDB)	0.00	0.05	0.03	0.04	0.15	0.15	0.06	0.15	-0.06	-0.01	0.17
	1.00	0.27	0.48	0.43	0.00	0.00	0.24	0.00	0.24	0.83	0.00
VolSite/ Shaper	-0.05	0.00	-0.02	-0.01	0.10	0.10	0.00	0.10	-0.11	-0.06	0.11
	0.27	1.00	0.70	0.75	0.01	0.02	0.93	0.02	0.02	0.18	0.00
Shaper (PDB)	-0.03	0.02	0.00	0.00	0.12	0.11	0.02	0.11	-0.09	-0.04	0.13
	0.48	0.70	1.00	0.95	0.00	0.01	0.64	0.01	0.05	0.34	0.00
Shaper	-0.04	0.01	0.00	0.00	0.11	0.11	0.02	0.11	-0.09	-0.05	0.13
	0.43	0.75	0.95	1.00	0.01	0.01	0.69	0.01	0.04	0.31	0.00
SiteAlign	-0.15	-0.10	-0.12	-0.11	0.00	0.00	-0.09	0.00	-0.21	-0.16	0.02
	0.00	0.01	0.00	0.01	1.00	0.93	0.02	0.93	0.00	0.00	0.64
SiteEngine	-0.15	-0.10	-0.11	-0.11	0.00	0.00	-0.09	0.00	-0.20	-0.16	0.02
	0.00	0.02	0.01	0.01	0.93	1.00	0.02	1.00	0.00	0.00	0.57
SiteHopper	-0.06	0.00	-0.02	-0.02	0.09	0.09	0.00	0.09	-0.11	-0.07	0.11
	0.24	0.93	0.64	0.69	0.02	0.02	1.00	0.03	0.01	0.15	0.01
SMAP	-0.15	-0.10	-0.11	-0.11	0.00	0.00	-0.09	0.00	-0.20	-0.16	0.02
	0.00	0.02	0.01	0.01	0.93	1.00	0.03	1.00	0.00	0.00	0.58
TIFP (PDB)	0.06	0.11	0.09	0.09	0.21	0.20	0.11	0.20	0.00	0.05	0.22
	0.24	0.02	0.05	0.04	0.00	0.00	0.01	0.00	1.00	0.33	0.00
TIFP	0.01	0.06	0.04	0.05	0.16	0.16	0.07	0.16	-0.05	0.00	0.18
	0.83	0.18	0.34	0.31	0.00	0.00	0.15	0.00	0.33	1.00	0.00
TM-align	-0.17	-0.11	-0.13	-0.13	-0.02	-0.02	-0.11	-0.02	-0.22	-0.18	0.00
	0.00	0.00	0.00	0.00	0.64	0.57	0.01	0.58	0.00	0.00	1.00

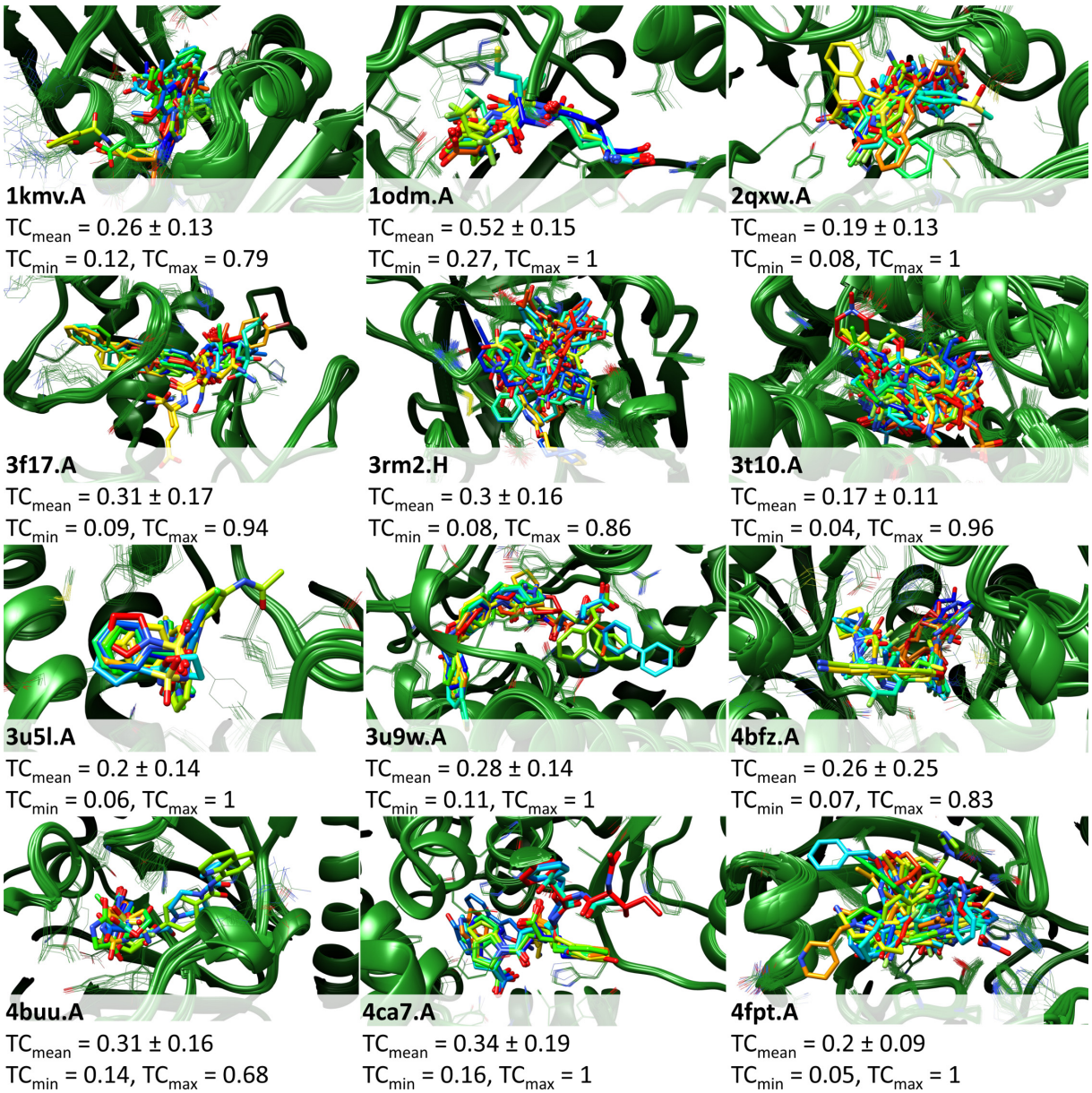
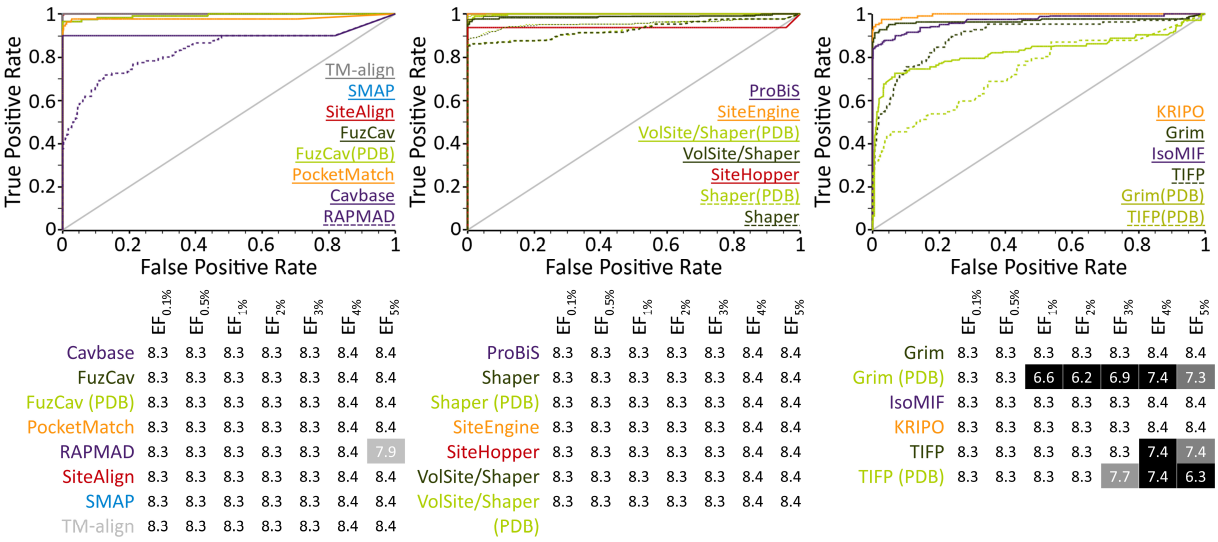


Figure A2.1. Binding site alignments for dataset 1. Binding site alignments and statistics of the Tanimoto coefficients for all pairs within the groups of structures with identical sequences are given. All figures were generated using UCSF Chimera[1].

Table A2.28. Statistics of the Tanimoto coefficients of dataset 1.2. The exclusion of highly dissimilar ligands within each group of dataset 1 led to this reduced dataset and the statistics of the pairwise Tanimoto coefficients of ligands within the groups are given.

PDB-ID.Chain	Mean	Standard Deviation	Minimum	Maximum
1kmv.A	0.79	0.00	0.79	0.79
1odm.A	0.72	0.09	0.63	1.00
2qxw.A	1.00	0.00	1.00	1.00
3f17.A	0.75	0.00	0.75	0.76
3rm2.H	0.71	0.06	0.63	0.86
3t10.A	0.73	0.11	0.65	0.96
3u5l.A	1.00	0.00	1.00	1.00
3u9w.A	0.67	0.02	0.64	0.69
4bfz.A	0.72	0.10	0.61	0.83
4buu.A	0.68	0.00	0.68	0.68
4ca7.A	0.86	0.11	0.79	1.00
4fpt.A	0.73	0.03	0.69	0.75



	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	8.3	8.3	8.3	8.3	8.3	8.4	8.4
FuzCav	8.3	8.3	8.3	8.3	8.3	8.4	8.4
FuzCav (PDB)	8.3	8.3	8.3	8.3	8.3	8.4	8.4
PocketMatch	8.3	8.3	8.3	8.3	8.3	8.4	8.4
RAPMAD	8.3	8.3	8.3	8.3	8.3	8.4	7.9
SiteAlign	8.3	8.3	8.3	8.3	8.3	8.4	8.4
SMAP	8.3	8.3	8.3	8.3	8.3	8.4	8.4
TM-align	8.3	8.3	8.3	8.3	8.3	8.4	8.4

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
ProBiS	8.3	8.3	8.3	8.3	8.3	8.4	8.4
Shaper	8.3	8.3	8.3	8.3	8.3	8.4	8.4
Shaper (PDB)	8.3	8.3	8.3	8.3	8.3	8.4	8.4
SiteEngine	8.3	8.3	8.3	8.3	8.3	8.4	8.4
SiteHopper	8.3	8.3	8.3	8.3	8.3	8.4	8.4
VolSite/Shaper	8.3	8.3	8.3	8.3	8.3	8.4	8.4
VolSite/Shaper (PDB)	8.3	8.3	8.3	8.3	8.3	8.4	8.4

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Grim	8.3	8.3	8.3	8.3	8.3	8.4	8.4
Grim (PDB)	8.3	8.3	6.6	6.2	6.9	7.4	7.3
IsoMIF	8.3	8.3	8.3	8.3	8.3	8.4	8.4
KRIPO	8.3	8.3	8.3	8.3	8.3	8.4	8.4
TIFP	8.3	8.3	8.3	8.3	8.3	7.4	7.4
TIFP (PDB)	8.3	8.3	8.3	8.3	7.7	7.4	6.3

Figure A2.2. Evaluation of different binding site comparison tools with respect to dataset 1.2. Top: The ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is coloured according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. The score RefTversky (color) led to the highest AUC values for Shaper, Shaper(PDB), VolSite/Shaper, and VolSite/Shaper(PDB). Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

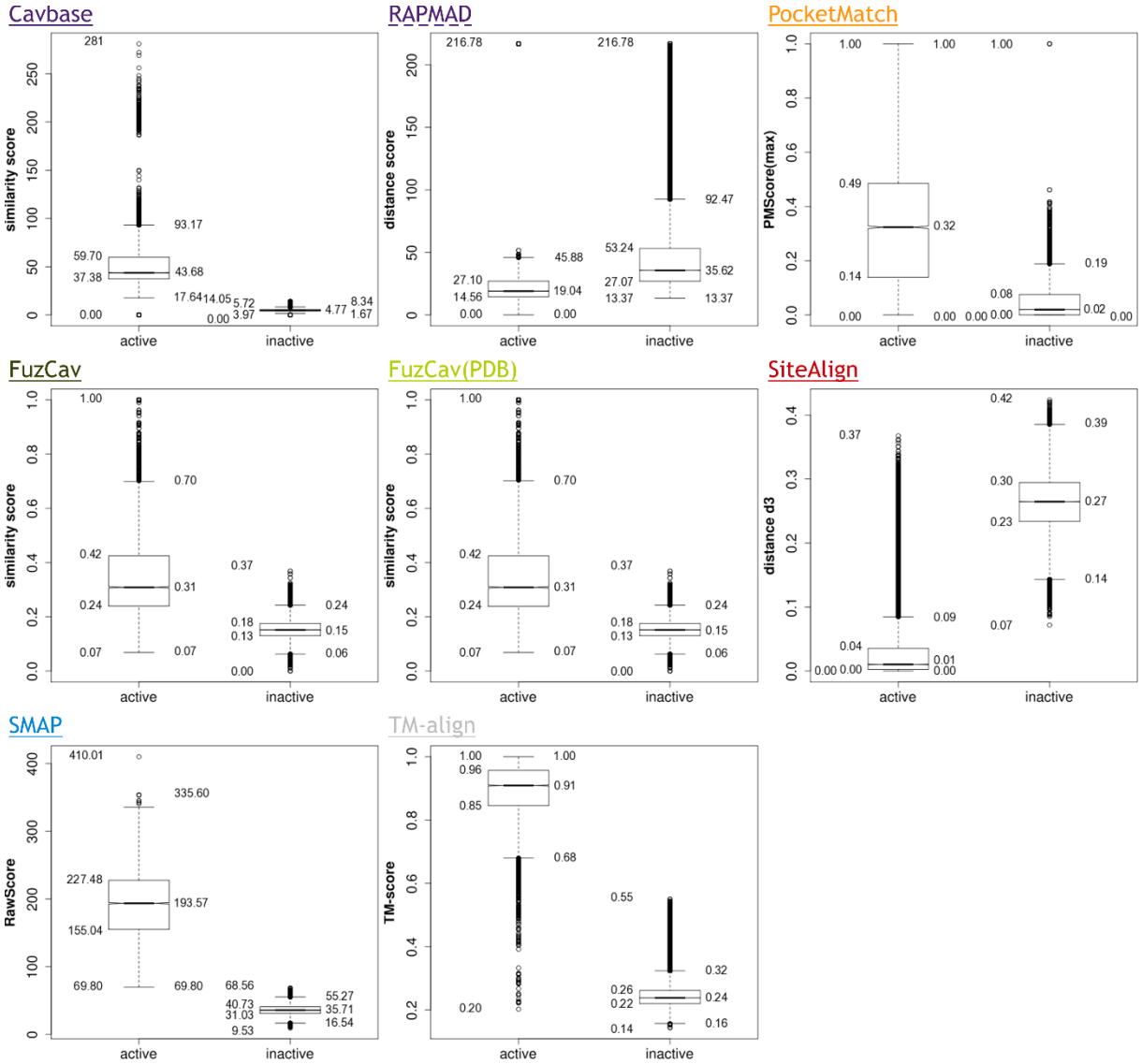


Figure A2.3. Score distributions of residue-based comparison methods. Boxplots to visualise the score distribution for similar ("active") and dissimilar ("inactive") pairs for the residue-based binding site comparison methods are given. The plots were created based on the dataset of structures with identical sequences. It has to be noted that the usage of RAPMAD and SiteAlign results in distances of the binding site pairs. The application of the remaining tools results in similarity scores. The values given in the plots correspond to the maximum, the upper whisker, the upper quartile, the median, the lower quartile, the lower whisker, and the minimum.

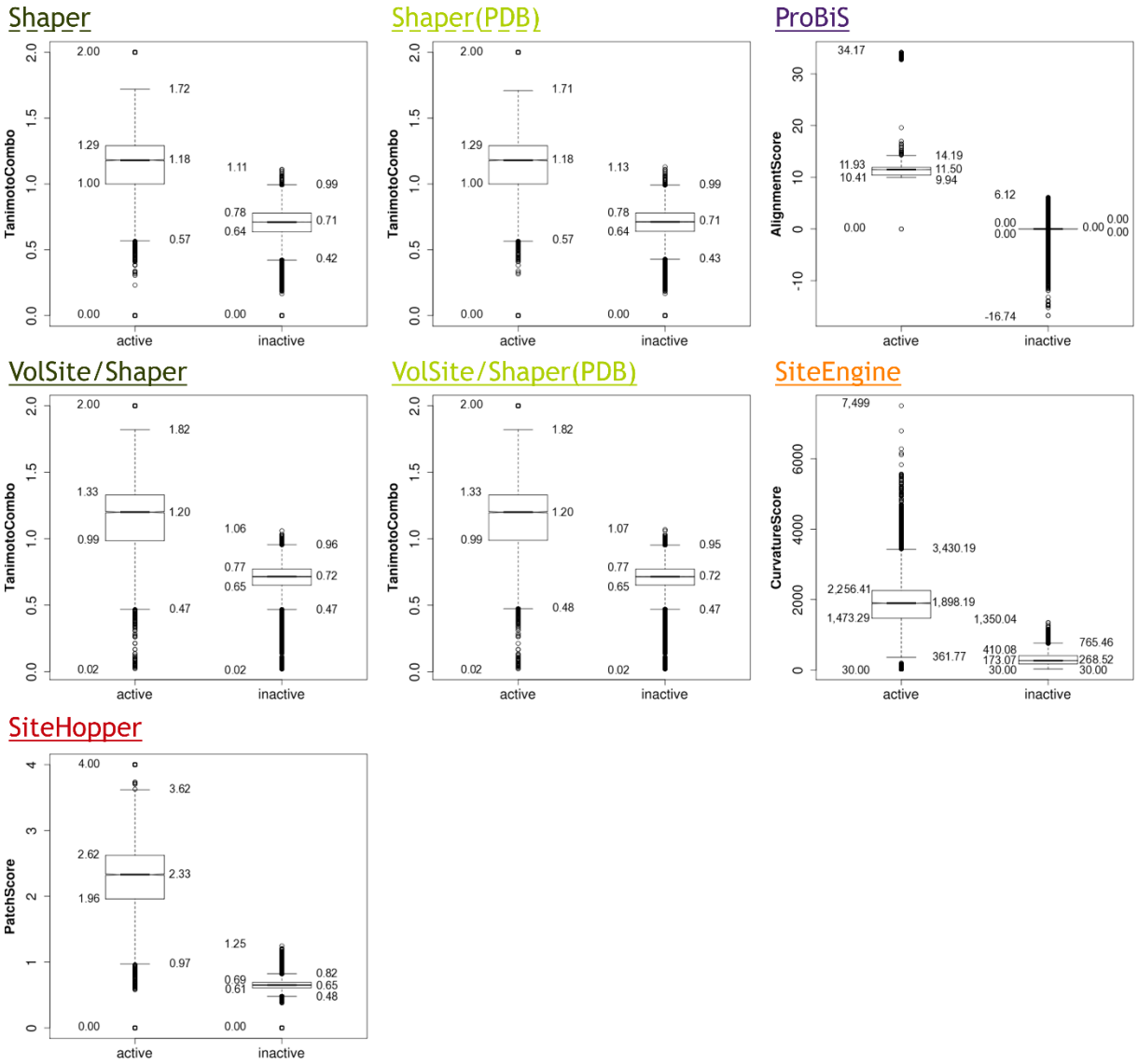


Figure A2.4. Score distributions of surface-based comparison methods. Boxplots to visualise the score distribution for similar ("active") and dissimilar ("inactive") pairs for the surface-based binding site comparison methods are given. The plots were created based on the dataset of structures with identical sequences. The values given in the plots correspond to the maximum, the upper whisker, the upper quartile, the median, the lower quartile, the lower whisker, and the minimum.

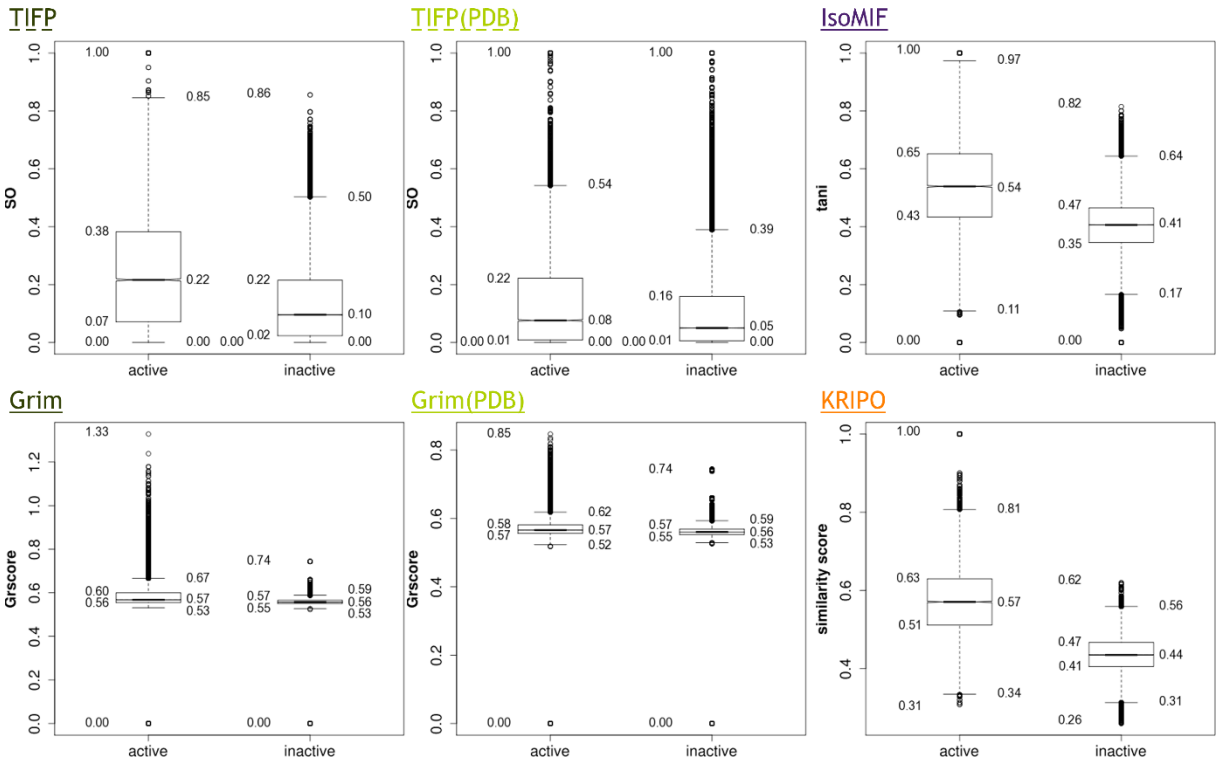


Figure A2.5. Score distributions of interaction-based comparison methods. Boxplots to visualise the score distribution for similar ("active") and dissimilar ("inactive") pairs for the interaction-based binding site comparison methods are given. The plots were created based on the dataset of structures with identical sequences. The values given in the plots correspond to the maximum, the upper whisker, the upper quartile, the median, the lower quartile, the lower whisker, and the minimum.

Table A2.29. Statistics of the pairwise binding site RMSD values in the structure groups of datasets 1 and 2.

PDB-ID.Chain	Property	Mean [Å]	Standard Deviation [Å]	Minimum [Å]	Maximum [Å]
Dataset 1					
1kmv.A	RMSD(Cα)	0.46	0.19	0.09	1.05
	RMSD(all atoms)	1.24	0.32	0.37	2.09
1odm.A	RMSD(Cα)	0.14	0.05	0.04	0.28
	RMSD(all atoms)	0.49	0.24	0.08	1.05
2qxw.A	RMSD(Cα)	0.55	0.22	0.09	1.16
	RMSD(all atoms)	0.85	0.27	0.32	1.41
3f17.A	RMSD(Cα)	0.34	0.11	0.13	0.71
	RMSD(all atoms)	0.72	0.19	0.26	1.13
3rm2.H	RMSD(Cα)	0.28	0.10	0.07	0.74
	RMSD(all atoms)	0.58	0.15	0.16	1.04
3t10.A	RMSD(Cα)	1.27	0.72	0.09	2.53
	RMSD(all atoms)	1.38	0.65	0.13	2.39
3u5l.A	RMSD(Cα)	0.25	0.10	0.06	0.47
	RMSD(all atoms)	0.44	0.13	0.11	0.67
3u9w.A	RMSD(Cα)	0.18	0.04	0.09	0.31
	RMSD(all atoms)	0.36	0.08	0.10	0.53
4bfz.A	RMSD(Cα)	0.82	0.32	0.15	1.44
	RMSD(all atoms)	1.18	0.32	0.54	1.72
4buu.A	RMSD(Cα)	0.59	0.50	0.08	1.44
	RMSD(all atoms)	1.08	0.57	0.26	2.03
4ca7.A	RMSD(Cα)	0.14	0.03	0.07	0.20
	RMSD(all atoms)	0.35	0.11	0.15	0.56
4fpt.A	RMSD(Cα)	0.27	0.08	0.04	0.51
	RMSD(all atoms)	0.61	0.12	0.13	0.91

Table A2.29 *(continued)*. Statistics of the pairwise binding site RMSD values in the structure groups of datasets 1 and 2.

PDB-ID.Chain	Property	Mean [Å]	Standard deviation [Å]	Minimum [Å]	Maximum [Å]
Dataset 2					
1cz2.A	RMSD(Cα)	1.21	0.33	0.01	1.97
	RMSD(all atoms)	1.76	0.40	0.02	2.50
1diu.A	RMSD(Cα)	0.14	0.06	0.03	0.28
	RMSD(all atoms)	0.35	0.10	0.10	0.57
1eio.A	RMSD(Cα)	1.21	0.19	0.87	1.50
	RMSD(all atoms)	1.61	0.28	1.11	1.98
1j5i.A	RMSD(Cα)	1.09	0.32	0.25	1.91
	RMSD(all atoms)	1.80	0.31	0.86	2.69
1kgl.A	RMSD(Cα)	0.41	0.10	0.20	0.70
	RMSD(all atoms)	0.80	0.16	0.38	1.16
1mux.A	RMSD(Cα)	11.55	2.51	4.04	19.08
	RMSD(all atoms)	11.46	2.45	4.43	19.12
1t84.A	RMSD(Cα)	2.56	0.93	0.70	4.42
	RMSD(all atoms)	3.24	0.89	1.12	5.02
1tvc.A	RMSD(Cα)	1.13	0.18	0.65	1.49
	RMSD(all atoms)	1.85	0.22	1.38	2.42
1yho.A	RMSD(Cα)	0.53	0.12	0.28	0.99
	RMSD(all atoms)	1.11	0.22	0.62	1.84
2jt2.A	RMSD(Cα)	0.19	0.05	0.11	0.37
	RMSD(all atoms)	0.52	0.08	0.27	0.74
2k31.A	RMSD(Cα)	0.31	0.06	0.15	0.47
	RMSD(all atoms)	1.00	0.17	0.54	1.48
2k5t.A	RMSD(Cα)	1.38	0.29	0.78	2.37
	RMSD(all atoms)	2.45	0.48	1.53	3.60
2l0x.A	RMSD(Cα)	1.17	0.47	0.31	2.25
	RMSD(all atoms)	1.94	0.61	0.74	3.43
2l2s.A	RMSD(Cα)	0.73	0.26	0.28	1.33
	RMSD(all atoms)	0.94	0.22	0.39	1.43
2l8r.A	RMSD(Cα)	0.94	0.17	0.53	1.47
	RMSD(all atoms)	1.47	0.19	0.94	2.12
2lzg.A	RMSD(Cα)	0.83	0.11	0.65	0.99
	RMSD(all atoms)	1.35	0.15	1.14	1.59
2z2d.A	RMSD(Cα)	1.34	0.27	0.77	2.10
	RMSD(all atoms)	1.96	0.36	1.24	3.00

Table A2.30. Statistics of the binding site descriptors for the NMR ensembles of dataset 2. The binding site descriptors were calculates with DoGSite[21]. Mean, standard deviation, minimum, and maximum are given for four descriptors of ligand-defined sites.

PDB-ID. Chain	Number of site atoms	Number of hydrogen bond acceptor atoms	Number of hydrogen bond donor atoms	Number of aromatic atoms
1cz2.A	184±27, 146, 238	26±5, 20, 34	21±5, 14, 34	6±2, 2, 10
1diu.A	275±29, 237, 334	43±5, 35, 55	55±5, 49, 66	30±2, 27, 35
1eio.A	281±29, 226, 304	47±7, 35, 55	47±6, 39, 56	54±6, 42, 61
1j5i.A	96±47, 59, 256	22±8, 14, 52	16±7, 10, 46	5±5, 0, 16
1kgl.A	185±21, 150, 233	27±3, 22, 34	26±4, 19, 34	32±3, 26, 37
1mux.A	70±34, 42, 205	9±7, 2, 37	5±5, 2, 24	10±3, 3, 24
1t84.A	163±36, 112, 247	26±7, 16, 44	24±5, 17, 35	22±4, 16, 29
1tvc.A	375±151, 251, 686	57±25, 34, 110	54±26, 32, 109	43±13, 22, 67
1yho.A	344±112, 179, 537	55±18, 25, 87	53±18, 25, 82	48±12, 26, 77
2jt2.A	235±37, 177, 337	41±6, 33, 61	44±7, 32, 63	44±8, 30, 58
2k31.A	142±9, 129, 161	19±2, 14, 23	23±2, 16, 27	23±2, 19, 26
2k5t.A	157±50, 69, 259	22±8, 9, 38	25±9, 10, 41	15±7, 4, 29
2l0x.A	98±40, 33, 165	19±8, 7, 33	19±8, 6, 32	5±3, 0, 11
2l2s.A	245±61, 123, 335	38±10, 18, 57	36±11, 14, 53	42±4, 32, 49
2l8r.A	199±21, 156, 230	32±5, 24, 39	34±4, 27, 43	6±2, 2, 10
2lzg.A	193±73, 108, 321	27±10, 14, 43	27±11, 15, 45	26±9, 14, 42
2z2d.A	171±54, 111, 333	28±10, 19, 58	25±8, 16, 47	29±10, 18, 57

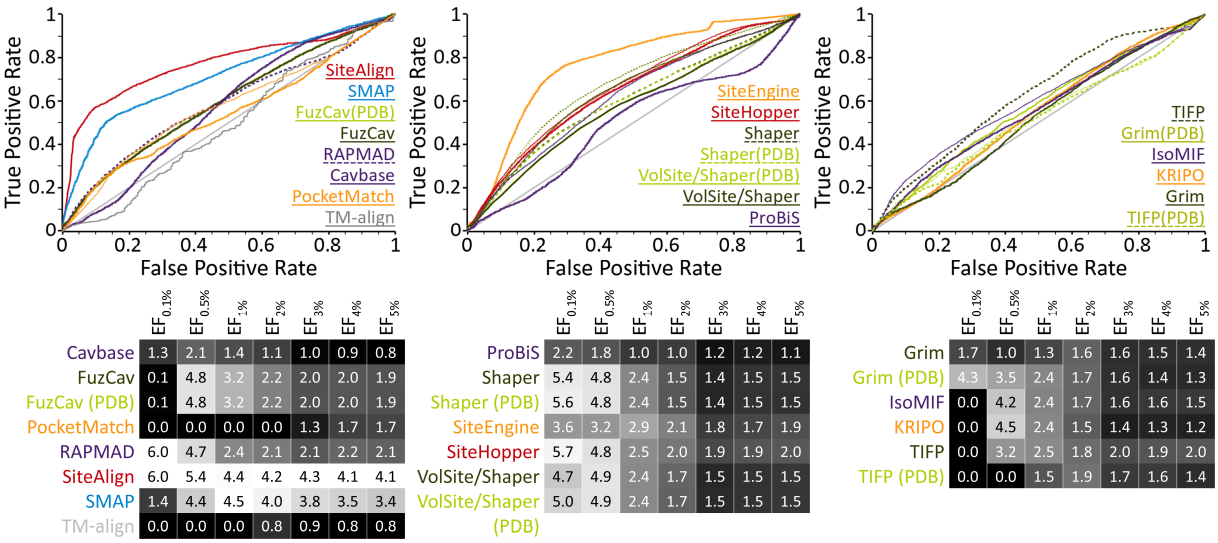


Figure A2.6. Evaluation of different binding site comparison tools with respect to dataset 3 (one, two, three, four, and five substitutions by physicochemically different residues). Top: The ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is colored according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. PocketMatch showed the best AUC for the score $PMScore_{min}$ (thin orange line). The scores SVA, RefTversky (color), RefTversky (color), RefTversky (color), RefTversky (color), and ColorTanimoto led to the highest AUC values for ProBiS, Shaper, Shaper(PDB), VolSite/Shaper, VolSite/Shaper(PDB), and SiteHopper, respectively (thin lines). The highest AUC was obtained for IsoMIF and TIFP(PDB) when using taniM and the Tanimoto coefficient, respectively (thin lines). Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

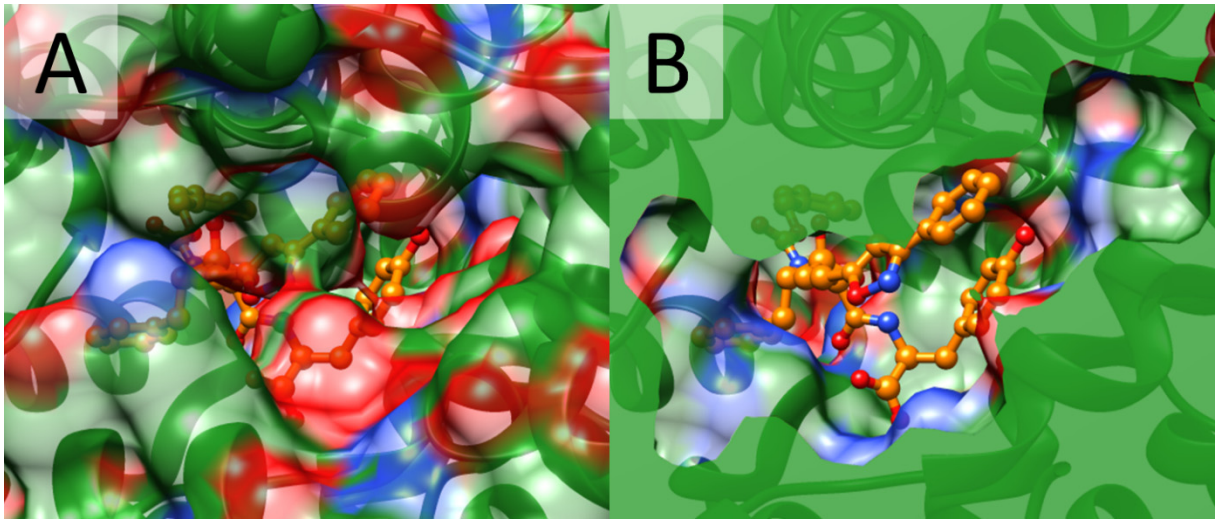
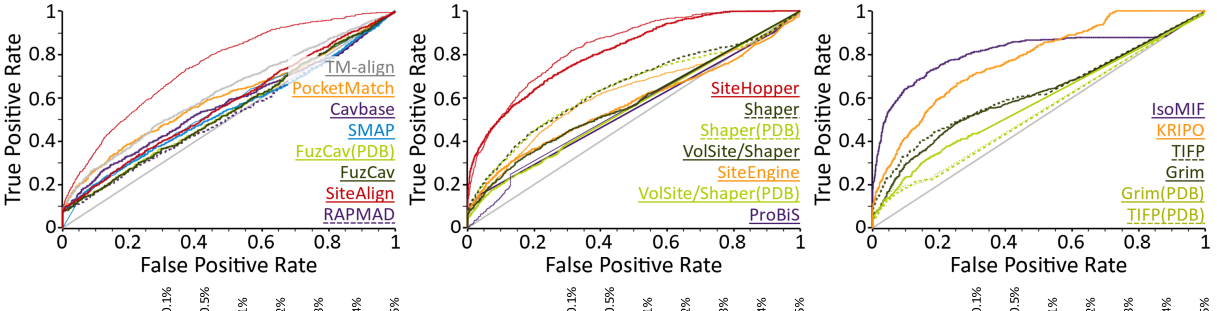


Figure A2.7. The buried binding site of angiotensin-converting enzyme. The ligand binding site of angiotensin-converting enzyme (PDB-ID 4ca7 chain A) is depicted in green in the surface representation (A) and the clipped surface representation (B). The ligand with the id 3EF is shown as an orange balls-and-sticks model. The figure was generated using UCSF Chimera[1].

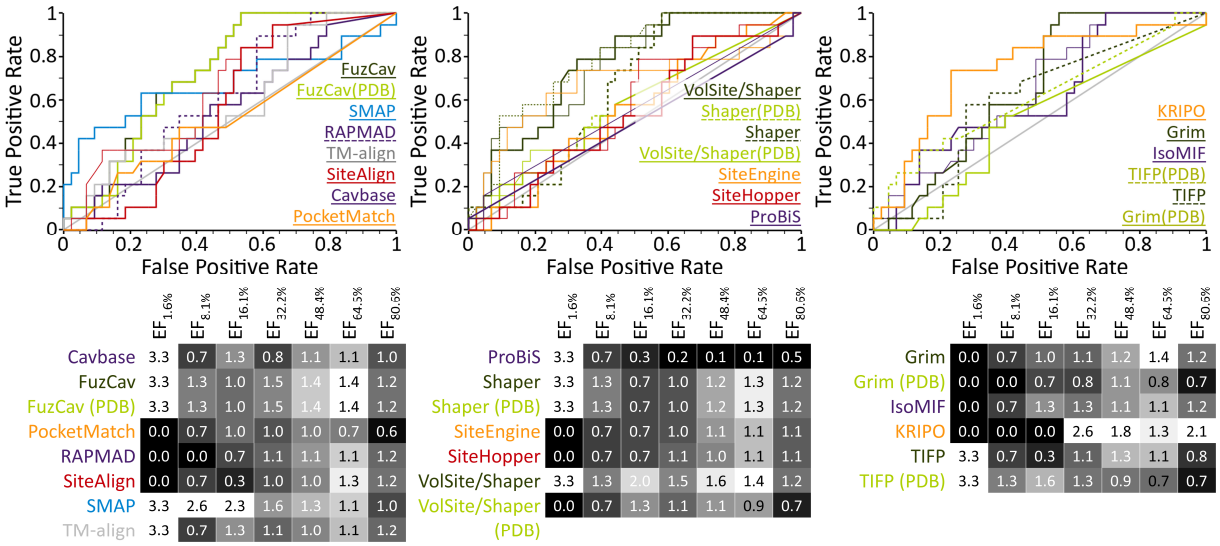


	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Cavbase	7.6	7.6	7.0	4.1	3.4	3.0	2.6
FuzCav	7.6	7.6	7.4	4.2	3.0	2.5	2.3
FuzCav (PDB)	7.6	7.6	7.4	4.2	3.1	2.5	2.3
PocketMatch	7.6	7.6	7.3	4.9	4.2	3.7	3.3
RAPMAD	7.6	7.6	7.3	4.4	3.1	2.4	2.1
SiteAlign	7.6	7.6	7.5	4.8	3.5	2.9	2.5
SMAP	7.6	7.6	7.4	4.7	3.4	2.9	2.6
TM-align	7.6	7.6	7.6	5.3	4.2	3.7	3.3

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
ProBiS	7.6	7.4	7.5	5.0	3.7	3.3	2.9
Shaper	7.6	7.6	7.3	5.2	4.3	3.7	3.3
Shaper (PDB)	7.6	7.6	7.3	5.1	4.3	3.7	3.4
SiteEngine	7.6	7.6	7.0	4.9	3.8	3.4	3.0
SiteHopper	7.6	7.6	7.6	7.0	6.6	6.4	5.8
VolSite/Shaper	7.6	7.6	6.4	4.2	3.3	2.9	2.6
VolSite/Shaper (PDB)	7.6	7.6	5.2	3.3	2.8	2.6	2.4

	$EF_{0.1\%}$	$EF_{0.5\%}$	$EF_{1\%}$	$EF_{2\%}$	$EF_{3\%}$	$EF_{4\%}$	$EF_{5\%}$
Grim	7.6	7.6	4.4	3.6	2.8	2.8	2.7
Grim (PDB)	0.0	4.4	4.9	3.7	2.5	2.6	2.6
IsoMIF	7.6	7.6	7.6	7.4	6.9	6.2	6.0
KRIPO	7.6	7.6	7.6	5.9	5.4	4.6	4.3
TIFP	7.6	7.6	7.0	5.5	4.6	4.0	3.7
TIFP (PDB)	0.0	5.5	4.2	2.8	2.7	2.4	2.3

Figure A2.8. Evaluation of different binding site comparison tools with respect to dataset 6.2. Top: ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is colored according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. The best AUC for SiteAlign resulted from the d1 distance (thin red line). For ProBiS, SiteEngine, and SiteHopper the similarity measures SVA, LowResolutionScore, and ShapeTanimoto yielded the best AUC values (thin lines). For TIFP(PDB), the use of the Hamming distance led to the best results with respect to AUC (thin line). Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.



	$EF_{1.6\%}$	$EF_{3.1\%}$	$EF_{16.1\%}$	$EF_{32.2\%}$	$EF_{48.4\%}$	$EF_{64.5\%}$	$EF_{80.5\%}$
Cavbase	3.3	0.7	1.3	0.8	1.1	1.1	1.0
FuzCav	3.3	1.3	1.0	1.5	1.4	1.4	1.2
FuzCav (PDB)	3.3	1.3	1.0	1.5	1.4	1.4	1.2
PocketMatch	0.0	0.7	1.0	1.0	1.0	0.7	0.6
RAPMAD	0.0	0.0	0.7	1.1	1.1	1.1	1.2
SiteAlign	0.0	0.7	0.3	1.0	1.0	1.3	1.2
SMAP	3.3	2.6	2.3	1.6	1.3	1.1	1.0
TM-align	3.3	0.7	1.3	1.1	1.0	1.1	1.2

	$EF_{1.6\%}$	$EF_{3.1\%}$	$EF_{16.1\%}$	$EF_{32.2\%}$	$EF_{48.4\%}$	$EF_{64.5\%}$	$EF_{80.5\%}$
ProBiS	3.3	0.7	0.3	0.2	0.1	0.1	0.5
Shaper	3.3	1.3	0.7	1.0	1.2	1.3	1.2
Shaper (PDB)	3.3	1.3	0.7	1.0	1.2	1.3	1.2
SiteEngine	0.0	0.7	0.7	1.0	1.1	1.1	1.1
SiteHopper	0.0	0.7	0.7	1.1	1.0	1.1	1.1
VolSite/Shaper	3.3	1.3	2.0	1.5	1.6	1.4	1.2
VolSite/Shaper (PDB)	0.0	0.7	1.3	1.1	1.1	0.9	0.7

	$EF_{1.6\%}$	$EF_{3.1\%}$	$EF_{16.1\%}$	$EF_{32.2\%}$	$EF_{48.4\%}$	$EF_{64.5\%}$	$EF_{80.5\%}$
Grim	0.0	0.7	1.0	1.1	1.2	1.4	1.2
Grim (PDB)	0.0	0.0	0.7	0.8	1.1	0.8	0.7
IsoMIF	0.0	0.7	1.3	1.3	1.1	1.1	1.2
KRIPO	0.0	0.0	0.0	2.6	1.8	1.3	2.1
TIFP	3.3	0.7	0.3	1.1	1.3	1.1	0.8
TIFP (PDB)	3.3	1.3	1.6	1.3	0.9	0.7	0.7

Figure A2.9. Evaluation of different binding site comparison tools with respect to dataset 7.2. Top: The ROC curves for residue- (left), surface- (centre), and interaction-based (right) comparison methods. The name of the tool is colored according to its corresponding ROC curve. The binding site comparison tools are sorted in descending order with respect to their AUC. The thin red line represents the resulting ROC curve for SiteAlign when using the distance d1. Thin lines represent the ROC curves for ProBiS, Shaper, Shaper(PDB), VolSite/Shaper, VolSite/Shaper(PDB), SiteEngine and SiteHopper when using the scoring schemes SVA, FitTversky (color), FitTversky (color), RefTversky (color), Tanimoto (fit), DistanceScore, and ShapeTanimoto, respectively. The thin line represents the resulting ROC curve for IsoMIF and the taniMW score. Bottom: EFs for residue-, surface-, and interaction-based comparison methods. A linear colour gradient ranging from white for the highest value to grey to black for the lowest value was applied for the EFs at different percentages of the screened dataset.

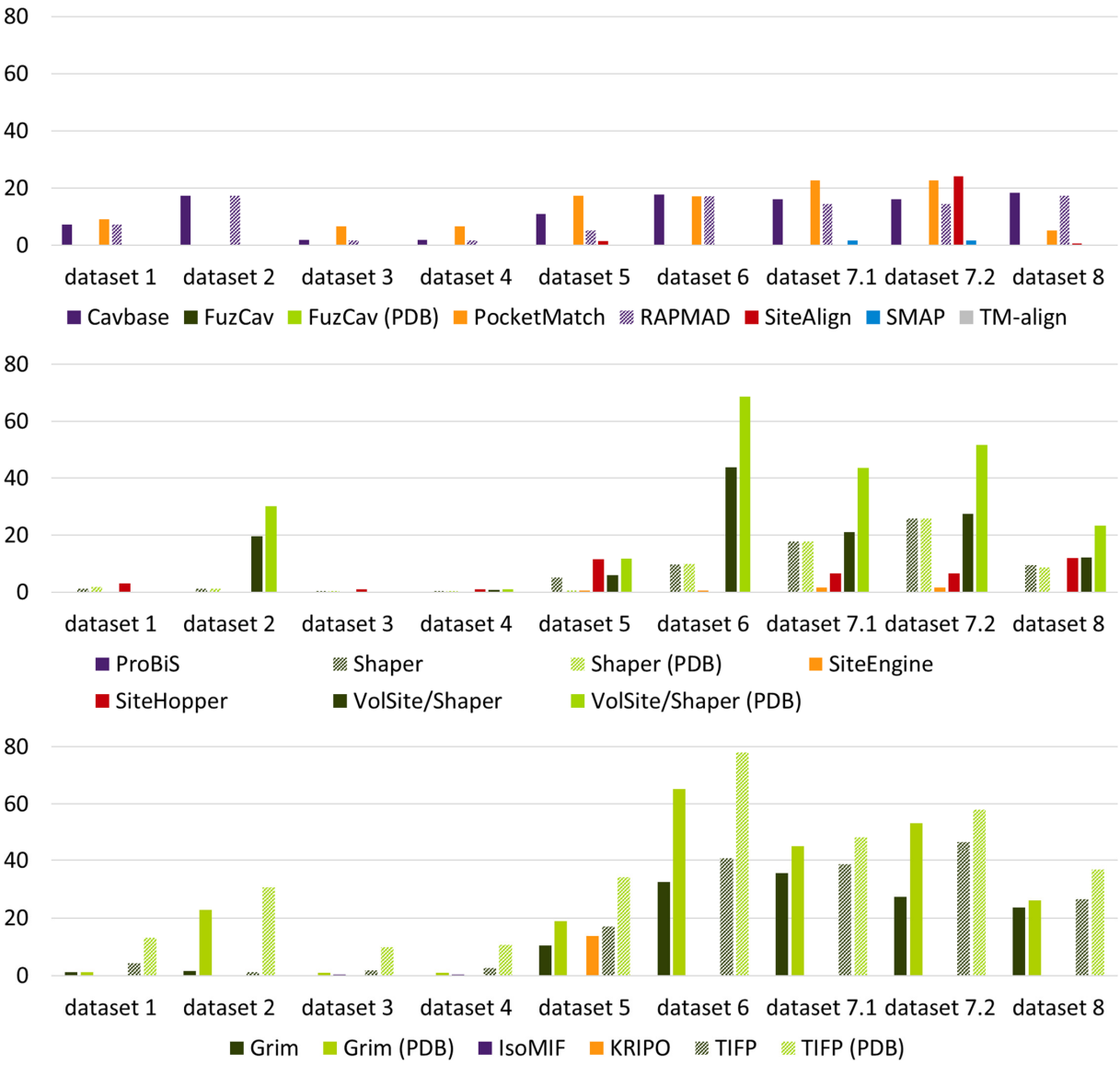


Figure A2.10. Failure rates of binding site comparison methods. The failure rates (percentages) of all residue-based (top), surface-based (middle), and interaction-based (bottom) binding site comparison methods are given for all datasets analysed in this study.

Table A2.31. Cut-off values for all comparison methods set defined using the Youden's J statistic[22]. The thresholds were defined separately for each dataset based on the corresponding ROC curves. Means and standard deviations for the cut-off values for all datasets are given together with the corresponding score and the score range. The distance measures of RAPMAD and SiteAlign were transformed to the corresponding similarities.

Method	Dataset 1	Dataset 1.2	Dataset 2	Dataset 3	Dataset 4	Dataset 6	Dataset 6.2
Cavbase	15.85	21.75	8.52	35.57	34.15	5.70	5.70
FuzCav (PDB)	0.21	0.28	0.22	0.27	0.36	0.21	0.21
FuzCav	0.21	0.28	0.22	0.27	0.36	0.21	0.22
Grim (PDB)	0.57	0.59	0.57	0.57	0.57	0.57	0.56
Grim	0.57	0.59	0.58	0.56	0.56	0.56	0.56
IsoMIF	0.49	0.62	0.50	0.60	0.62	0.37	0.35
KRIPO	0.50	0.52	0.51	0.56	0.56	0.46	0.44
PocketMatch	0.16	0.17	0.29	0.42	0.40	0.11	0.11
ProBiS	8.03	7.96	8.66	11.42	11.34	0.01	0.01
RAPMAD	0.89	0.8	0.75	0.74	0.73	0.90	0.95
VolSite/ Shaper (PDB)	0.89	0.95	0.88	1.20	1.15	0.75	0.70
VolSite/ Shaper	0.89	0.95	0.89	1.20	1.15	0.73	0.71
Shaper (PDB)	0.89	0.97	0.90	1.17	1.12	0.71	0.71
Shaper	0.89	0.97	0.90	1.16	1.12	0.74	0.74
SiteAlign	0.85	0.85	0.85	0.97	0.97	0.76	0.77
SiteEngine	875.25	1,444.60	244.00	1,431.71	1,516.61	417.60	417.60
SiteHopper	0.95	1.57	1.11	2.23	2.21	0.64	0.65
SMAP	69.18	98.17	92.21	184.71	218.71	44.83	51.47
TIFP (PDB)	0.20	0.38	0.14	0.10	0.05	0.08	0.03
TIFP	0.23	0.20	0.20	0.05	0.03	0.13	0.09
TM-align	0.49	0.64	0.67	0.85	0.85	0.25	0.25

Table A2.31 *(continued)*. Cut-off values for all comparison methods set defined using the Youden's J statistic[22]. The thresholds were defined separately for each dataset based on the corresponding ROC curves. Means and standard deviations for the cut-off values for all datasets are given together with the corresponding score and the score range. The distance measures of RAPMAD and SiteAlign were transformed to the corresponding similarities.

Method	Dataset 7	Dataset 7.2	Dataset 8	Mean	Standard Deviation	Score	Score Range
Cavbase	1.34	1.34	8.33	13.83	12.71	similarity score	n/a
FuzCav (PDB)	0.10	0.10	0.20	0.22	0.08	similarity score	[0,1]
FuzCav	0.10	0.10	0.20	0.22	0.08	similarity score	[0,1]
Grim (PDB)	0.55	0.55	0.59	0.57	0.01	Grscore	n/a
Grim	0.27	0.55	0.58	0.54	0.10	Grscore	n/a
IsoMIF	0.36	0.36	0.46	0.47	0.11	tani	[0,1]
KRIPO	0.44	0.44	0.56	0.50	0.05	similarity score	[0,1]
PocketMatch	0.03	0.03	0.26	0.20	0.14	$PMScore_{max}$	[0,1]
ProBiS	2.70	2.70	2.98	5.58	4.39	Alignment Score	n/a
RAPMAD	0.81	0.81	0.97	0.83	0.09	1-distance	[0,1]
VolSite/ Shaper (PDB)	0.80	0.22	0.98	0.85	0.27	Tanimoto (combo)	[0,2]
VolSite/ Shaper	0.71	0.70	0.99	0.89	0.18	Tanimoto (combo)	[0,2]
Shaper (PDB)	0.40	0.38	0.94	0.82	0.27	Tanimoto (combo)	[0,2]
Shaper	0.40	0.38	0.98	0.83	0.27	Tanimoto (combo)	[0,2]
SiteAlign	0.69	0.67	0.80	0.82	0.10	1-d3	[0,1]
SiteEngine	164.64	164.64	1,063.70	774.03	558.32	Curvature-Score	n/a
SiteHopper	0.65	0.65	0.87	1.15	0.63	PatchScore	[0,4]
SMAP	38.27	38.27	67.61	90.34	62.73	RawScore	n/a
TIFP (PDB)	0.11	0.07	0.35	0.15	0.12	Soergel	[0,1]
TIFP	0.15	0.15	0.17	0.14	0.07	Soergel	[0,1]
TM-align	0.20	0.20	0.38	0.48	0.26	TM-score	[0,1]

REFERENCES

[1] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.

[2] Willmann, D.; Lim, S.; Wetzel, S.; Metzger, E.; Jandausch, A.; Wilk, W.; Jung, M.; Forne, I.; Imhof, A.; Janzer, A.; Kirfel, J.; Waldmann, H.; Schule, R.; Buettner, R. Impairment of prostate cancer cell growth by a selective and reversible lysine-specific demethylase 1 inhibitor. *Int. J. Cancer* **2012**, *131* (11), 2704–2709.

[3] Haupt, V. J.; Daminelli, S.; Schroeder, M. Drug promiscuity in PDB: Protein binding site similarity is key. *PLoS ONE* **2013**, *8* (6), e65894.

[4] Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: New pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, *47* (3), 550–557.

[5] Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: Repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, *5* (7), e1000423.

[6] Yang, Y.; Li, G.; Zhao, D.; Yu, H.; Zheng, X.; Peng, X.; Zhang, X.; Fu, T.; Hu, X.; Niu, M.; Ji, X.; Zou, L.; Wang, J. Computational discovery and experimental verification of tyrosine kinase inhibitor pazopanib for the reversal of memory and cognitive deficits in rat model neurodegeneration. *Chem. Sci.* **2015**, *6* (5), 2812–2821.

[7] Niu, M.; Hu, J.; Wu, S.; Zhang, X.; Xu, H.; Zhang, Y.; Zhang, J.; Yang, Y. Structural bioinformatics-based identification of EGFR inhibitor gefitinib as a putative lead compound for BACE. *Chem. Biol. Drug Des.* **2014**, *83* (1), 81–88.

[8] Xie, L.; Evangelidis, T.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: Weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS Comput. Biol.* **2011**, *7* (4), e1002037.

[9] Yang, L.; Chen, J.; Shi, L.; Hudock, M. P.; Wang, K.; He, L. Identifying unexpected therapeutic targets via chemical-protein interactome. *PLoS ONE* **2010**, *5* (3), e9568.

[10] Defranchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS ONE* **2010**, *5* (8), e12214.

[11] Kinnings, S. L.; Jackson, R. M. Binding site similarity analysis for the functional classification of the protein kinase family. *J. Chem. Inf. Model.* **2009**, *49* (2), 318–329.

[12] Cleves, A. E.; Jain, A. N. Chemical and protein structural basis for biological crosstalk between PPARα and COX enzymes. *J. Comput.-Aided Mol. Des.* **2015**, *29* (2), 101–112.

[13] Xie, L.; Wang, J.; Bourne, P. E. *In silico* elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.* **2007**, *3* (11), e217.

[14] Babu, M.; Beloglazova, N.; Flick, R.; Graham, C.; Skarina, T.; Nocek, B.; Gagarinova, A.; Pogoutse, O.; Brown, G.; Binkowski, T. A.; Phanse, S.; Joachimiak, A.; Koonin, E. V.; Savchenko, A.; Emili, A.; Greenblatt, J.; Edwards, A. M.; Yakunin, A. F. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol. Microbiol.* **2011**, *79* (2), 484–502.

[15] Kinoshita, K.; Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **2003**, *12* (8), 1589–1595.

[16] Anand, P.; Sankaran, S.; Mukherjee, S.; Yeturu, K.; Laskowski, R. A.; Bhardwaj, A.; Bhagavat, R.; Brahmachari, S. K.; Chandra, N. Structural annotation of Mycobacterium tuberculosis proteome. *PLoS ONE* **2011**, *6* (10), e27044.

[17] Kuhn, D.; Weskamp, N.; Schmitt, S.; Hüllermeier, E.; Klebe, G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **2006**, *359* (4), 1023–1044.
[18] Ausiello, G.; Peluso, D.; Via, A.; Helmer-Citterich, M. Local comparison of protein structures highlights cases of convergent evolution in analogous functional sites. *BMC Bioinf.* **2007**, *8 Suppl 1*, S24.
[19] Tseng, Y. Y.; Li, W.-H. Classification of protein functional surfaces using structural characteristics. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (4), 1170–1175.
[20] DeLong, E. R.; DeLong, D. M.; Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44* (3), 837–845.
[21] Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.* **2012**, *52* (2), 360–372.
[22] Youden, W. J. Index for rating diagnostic tests. *Cancer* **1950**, *3* (1), 32–35.

APPENDIX 3: Secondary Structure Assignment

List A3.1. Residues which are not considered during the assignment of turns and secondary structures elements. The residues were found in the PDB as part of the protein macromolecule, but are characterised by atypical backbone structures.

00E, 00I, 00S, 010, 011, 01B, 02G, 02J, 02N, 03O, 04D, 05W, 0A9, 0AD, 0DQ, 0HQ, 0JT, 0JU, 0MG, 0OB, 0QE, 0QZ, 0R8, 0TJ, 0UH, 0W5, 0W6, 0XM, 0XN, 0YA, 10C, 11Q, 125, 126, 127, 12A, 175, 18M, 18Q, 1AP, 1BO, 1C3, 1CC, 1FC, 1HB, 1HD, 1KC, 1LU, 1MA, 1MG, 1OL, 1QQ, 1RN, 1SC, 1U8, 1VR, 1ZN, 22Q, 22W, 23G, 24M, 24O, 28H, 28K, 2A1, 2AR, 2AT, 2AU, 2BD, 2BT, 2BU, 2DA, 2DT, 2EG, 2GT, 2JF, 2JG, 2JV, 2KT, 2MA, 2MG, 2MU, 2N2, 2NT, 2OP, 2OT, 2PP, 2PR, 2SG, 2ST, 2UC, 2UE, 2X0, 32L, 34H, 39Y, 3A5, 3AZ, 3BY, 3CN, 3DA, 3FB, 3ME, 3PA, 3V2, 3V3, 3V7, 40A, 40C, 40G, 40T, 47C, 48V, 4AR, 4BA, 4CG, 4F3, 4FU, 4G6, 4H0, 4KY, 4L0, 4L8, 4LT, 4M9, 4MM, 4N7, 4N8, 4N9, 4NT, 4NU, 4OC, 4PC, 4SC, 4SU, 54L, 56J, 5AA, 5AT, 5BU, 5CG, 5CM, 5FC, 5FU, 5HC, 5HT, 5HU, 5IC, 5IU, 5MC, 5MU, 5NC, 5PC, 5PY, 5R0, 5R5, 5SE, 5SQ, 5UA, 5VV, 5XU, 5ZA, 60H, 62H, 63G, 63H, 64P, 64T, 66N, 68Z, 69P, 6E4, 6F5, 6FH, 6HA, 6HB, 6HC, 6HG, 6HT, 6IA, 6J9, 6L3, 6L9, 6MA, 6MZ, 6NA, 6OG, 6PO, 6RK, 6V9, 6VA, 6VF, 6VO, 707, 70U, 7AT, 7BB, 7DA, 7GA, 7GU, 7MG, 8AG, 8AN, 8BA, 8BB, 8FG, 8LR, 8MC, 8MG, 8OG, 92B, 9AT, 9BB, 9DK, 9DZ, 9GE, 9PR, 9TS, 9V7, A23, A2L, A2M, A38, A3A, A3P, A40, A43, A44, A47, A5L, A5M, A5O, A5R, A6A, A6C, A6G, A6U, A7E, A9Z, AA1, AAR, ABN, ABR, ABS, ABU, ACA, ACE, ACT, ACY, AD2, AEA, AF2, AFC, AG2, AH0, AHH, AJE, AKK, AKR, AKZ, ALQ, ALT, ALV, AMP, AMU, AMV, ANZ, AP7, APK, APN, AR0, AR7, ARF, AS, ASA, ASJ, ATD, ATL, ATM, AVC, AY0, AYE, AZ1, AZI, AZK, B27, B2A, B2F, B2I, B2N, B2V, B3A, B3D, B3E, B3K, B3L, B3M, B3Q, B3S, B3T, B3X, B3Y, B7C, BAL, BBC, BBS, BCX, BE2, BEZ, BGC, BGM, BIL, BLE, BMN, BNO, BOC, BOE, BOR, BP4, BPE, BRU, BTN, BUA, BZG, C12, C2L, C2S, C31, C34, C36, C37, C38, C42, C43, C45, C46, C49, C4S, C5L, C6G, C99, CA1, CAR, CBR, CBV, CCC, CCY, CDE, CDW, CEV, CF0, CFD, CFL, CFT, CFY, CFZ, CG1, CGN, CH, CH6, CH7, CHS, CJO, CKC, CLR, CLV, CM0, CMR, CMT, COI, CP1, CPN, CQ1, CQ2, CR0, CR2, CR5, CR7, CR8, CRF, CRG, CRK, CRO, CRQ, CRU, CRX, CSH, CSL, CTG, CWR, CX2, CXP, CY3, CYF, CZO, D00, DA, DAO, DBH, DC, DCL, DDG, DFI, DFO, DFT, DG, DG8, DGP, DHL, DIP, DIX, DIY, DKA, DM0, DMG, DNR, DOA, DOC, DT, DUZ, DYG, DYJ, DZM, E, E1H, E1X, ECC, ECQ, EDA, EFG, EHG, EIT, END, ESD, ETA, EXB, EXC, EYG, F3H, F4H, FA2, FAX, FBE, FBP, FDG, FDL, FE3, FHU, FMG, FMT, FMU, FOR, FOX, FPA, FPR, FRD, FUC, FUL, FUM, G2L, G2S, G31, G36, G38, G3A, G46, G47, G48, G49, G7M, GAL, GAO, GAU, GCK, GDO, GDP, GF2, GFL, GG7, GIC, GL3, GLC, GLK, GLZ, GM8, GMA, GMP, GMS, GMU, GN7, GNE, GNG, GOA, GPN, GRB, GS, GSR, GSS, GTP, GVE, GX1, GYC, GYS, H2U, HAO, HAQ, HCI, HEU, HF2, HG7, HIA, HM7, HM8, HM9, HMB, HMR, HN0, HN1, HOA, HPH, HR7, HS9, HSL, HSO, HSV, HT7, HY1, IBU, IC, IEY, IG, IGU, IIC, IL0, ILM, IMC, IP8, IPI, IU, IVA, JDT, JG3, KAC, KAG, KCQ, KI2, KPN, KWS, L2O, L3O, LAC, LAL, LCA, LCC, LCG, LEN, LGP, LHO, LHV, LIG, LKC, LML, LOV, LPD, LPL, LTA, LYJ, LYK, LYN, LYT, M1G, M2G, M5M, M9P, MA6, MA7, MAN, MAZ, MBN, MCM, MCR, MCY, MDF, MDO, MDP, ME6, MFC, MFD, MG1, MH6, MH9, MHE, MHW, MIA, MKE, MLI, MLL, MMT, MN1, MN2, MN7, MN8, MNU, MPR, MPT, MRG, MSU, MTR, MTU, MUB, MX3, MX4, MX5, MY1, MY2, MY3, MY5, MYR, N2G, N5M, N6G, N79, N7P, N80, NAG, NAK, NDG, NEH, NFA, NH2, NHH, NIT, NLO, NLW, NME, NMI, NMS, NMT, NOR, NRP, NRQ, NTA, NYG, NZH, O2G, OAR, OCE, ODA, ODR, OFM, OGX, OHU, OIL, OIM, OMC, OMG, OMU, ONE, OPR, OSL, OTT, OUD, OUE, OUH, OUI, OUK, OUR, P, P2T, P2U, P5P, PBE, PCS, PDU, PDW, PEA, PFX, PGA, PGN, PGP, PHA, PHL, PHQ, PIA, PIP, PIV, PJE, PJJ, PLF, PLJ, PLW, PN2, POL, PPI, PPU, PPW, PR3, PR4, PR7, PR9, PRN, PRQ, PRW, PSA, PST, PSU, PTL, PU, PVA, PVO, PVX, PXZ, PYO, PYR, QAC, QBT, QFG, QLG, QSC, QUA, QUO, R, RBD, RC7, RDG, RGL, RIA, RMP, RNG, RPC, RSQ, RTY, RUS, S2M, S4A, S4C, S4G, S6G, SC, SDE, SDG, SDH, SEL, SET, SIC, SIN, SMP, SMT, SNN, SPT, SRA, SSU, STA, SUI, SUJ, SUR, SWG, T23, T2S, T32, T38, T39, T3P, T41, T48, T49, T4S, T5O, T5S, T6A, TA2, TA3, TA4, TAF, TC1, TCP, TCY, TDY, TED, TEE, TFE, TFO, TFT, TGP, THO, TKL, TLB, TLC, TLN, TLX, TM2, TP1, TPC, TPL, TPN, TRW, TSP, TT, TTD, TTI, TTM, TYB, TYC, TYE, TYF, TYW, TYZ, U2L, U31, U33, U34, U36, U37, U8U, UAR, UBB, UBI, UCL, UD5, UDP, UF2, UFR, UFT, UMP, UMS, UMX, UPE, UPS, UPV, UR3, URE, URX, US1, US2, US3, US5, USC, USM, UVX, UZR, V9C, VAD, VAI, VLM, VLT, VME, VOL, W6Q, WCR, X, X9Q, XAD, XAL, XCL, XCP, XCR, XCT, XCY, XGL, XGR, XGU, XPB, XPC, XSN, XTF, XTH, XTL, XTR, XUA, XUG, XXY, XY1, XYG, XZA, Y, YAC, YCO, YG, YYG, Z, ZAD, ZBC, ZBU, ZCY, ZDU, ZGL, ZGU, ZHP, ZSC, ZSN, ZTH, ZZJ

List A3.2. Residues which are excluded from the backbone hydrogen atom placement. The backbone nitrogen atom of these residues has no open valencies.

02V, 05N, 0EA, 0FL, 0LF, 0Y8, 0YG, 11Q, 13E, 1JM, 22G, 2JC, 2MT, 33X, 34E, 36D, 3A0, 3A5, 3PX, 3V8, 45F, 45W, 4FB, 4N3, 5JP, 5PG, 66E, 73N, 73P, 9KK, 9WV, AET, AME, AUX, AYA, BMT, CNT, CPI, CSI, CXM, DAM, DC0, DI8, DLS, DMT, DNG, DNM, DPL, DPR, DSE, EME, FC0, FME, FP9, FVA, GNC, H5M, HGM, HY3, HYP, HZP, I4G, IML, IPG, LAV, LAY, LHE, LNT, MAA, MDH, MEA, MGG, MH6, MHL, MHU, MHV, MLE, MLU, MME, MMO, MNL, MOZ, MP8, MPQ, MSA, MV9, MVA, N2C, NCB, NCY, NLG, NLY, NMC, NZC, OIC, OLD, OTH, OZT, PCA, PDF, PH6, POM, PRJ, PRS, PUK, PXU, QIL, QPH, SAC, SAR, SC2, SL5, THC, TIG, TMD, TVA, TYJ, UMA, WLU, XPR, XXA, YCP, YNM, YYA, ZAE, ZYJ, ZYK

NORMAL 3 TURNS

Table A3.1. Frequencies and conformational parameters of residues within different types and at specific positions of normal 3 turns. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: n3-1								
Ala	384	0.81	99	0.62	184	1.16	101	0.63
Cys	70	1.01	28	1.21	17	0.73	25	1.08
Asp	419	1.27	90	0.82	217	1.97	112	1.02
Glu	230	0.62	108	0.88	70	0.57	52	0.42
Phe	272	1.21	97	1.29	54	0.72	121	1.61
Gly	341	0.82	159	1.14	89	0.64	93	0.67
His	169	1.28	48	1.09	45	1.02	76	1.73
Ile	324	1.02	141	1.33	7	0.07	176	1.66
Lys	199	0.65	78	0.76	61	0.60	60	0.59
Leu	504	0.98	221	1.29	123	0.72	160	0.93
Met	82	0.86	28	0.88	29	0.91	25	0.79
Asn	342	1.48	73	0.95	182	2.36	87	1.13
Pro	617	2.39	116	1.35	501	5.81	0	0.00
Gln	153	0.75	53	0.78	55	0.81	45	0.66
Arg	191	0.67	78	0.82	55	0.58	58	0.61
Ser	263	0.81	87	0.81	63	0.58	113	1.05
Thr	273	0.91	66	0.66	29	0.29	178	1.78
Val	364	0.92	152	1.15	7	0.05	205	1.55
Trp	116	1.48	39	1.49	21	0.80	56	2.14
Tyr	227	1.17	83	1.28	39	0.60	105	1.62
Other	16	0.58	8	0.87	4	0.44	4	0.44
type: n3-2								
Ala	53	0.62	30	1.05	9	0.31	14	0.49
Cys	10	0.80	4	0.96	2	0.48	4	0.96
Asp	60	1.01	18	0.91	19	0.96	23	1.16
Glu	52	0.78	15	0.68	21	0.95	16	0.72
Phe	37	0.91	12	0.88	9	0.66	16	1.18
Gly	131	1.74	35	1.39	76	3.03	20	0.80
His	32	1.35	10	1.26	16	2.02	6	0.76
Ile	42	0.73	9	0.47	18	0.94	15	0.78
Lys	48	0.87	13	0.71	19	1.03	16	0.87
Leu	56	0.60	23	0.74	13	0.42	20	0.65
Met	21	1.22	8	1.39	3	0.52	10	1.74
Asn	38	0.91	12	0.86	10	0.72	16	1.15
Pro	40	0.86	40	2.57	0	0.00	0	0.00
Gln	29	0.79	10	0.82	10	0.82	9	0.73
Arg	58	1.13	15	0.88	23	1.35	20	1.17
Ser	66	1.13	16	0.82	16	0.82	34	1.75
Thr	76	1.41	16	0.89	13	0.72	47	2.61
Val	67	0.94	22	0.92	21	0.88	24	1.01
Trp	26	1.84	8	1.69	11	2.33	7	1.48
Tyr	58	1.65	17	1.45	25	2.14	16	1.37
Other	2	0.40	1	0.61	0	0.00	1	0.61

NORMAL 4 TURNS

Table A3.2. Frequencies and conformational parameters of residues within different types and at specific positions of normal 4 turns. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*		*position 4*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: n4-1										
Ala	27,224	1.13	6,172	1.02	8,405	1.39	7,621	1.26	5,026	0.83
Cys	3,091	0.88	878	0.99	581	0.66	669	0.76	963	1.09
Asp	21,358	1.28	6,336	1.51	4,123	0.98	6,894	1.64	4,005	0.96
Glu	20,825	1.11	4,024	0.86	6,514	1.39	6,589	1.41	3,698	0.79
Phe	10,345	0.90	2,886	1.01	1,988	0.69	2,308	0.81	3,163	1.11
Gly	19,250	0.91	3,918	0.74	2,468	0.47	3,049	0.58	9,815	1.85
His	7,459	1.11	2,011	1.20	1,349	0.81	1,995	1.19	2,104	1.26
Ile	10,816	0.67	3,630	0.90	2,667	0.66	1,315	0.33	3,204	0.79
Lys	17,303	1.11	3,405	0.88	4,858	1.25	4,868	1.25	4,172	1.07
Leu	28,129	1.08	7,652	1.17	6,704	1.03	6,396	0.98	7,377	1.13
Met	4,946	1.02	1,217	1.00	1,207	1.00	1,261	1.04	1,261	1.04
Asn	14,071	1.20	3,650	1.24	2,126	0.72	4,595	1.56	3,700	1.26
Pro	12,532	0.96	4,338	1.32	7,103	2.16	1,091	0.33	0	0.00
Gln	11,707	1.13	2,143	0.83	2,745	1.06	3,629	1.40	3,190	1.23
Arg	14,992	1.04	3,287	0.91	4,034	1.12	3,995	1.11	3,676	1.02
Ser	19,820	1.21	4,288	1.04	5,142	1.25	6,106	1.49	4,284	1.04
Thr	11,978	0.79	3,127	0.82	2,358	0.62	3,122	0.82	3,371	0.89
Val	11,614	0.58	3,669	0.73	2,884	0.57	1,503	0.30	3,558	0.71
Trp	4,291	1.08	1,175	1.18	1,143	1.15	1,001	1.00	972	0.97
Tyr	8,895	0.90	2,365	0.96	1,757	0.71	2,172	0.88	2,601	1.05
Other	1,346	0.97	327	0.94	342	0.98	319	0.92	358	1.03
type: n4-2										
Ala	1,944	0.81	498	0.83	682	1.14	113	0.19	651	1.09
Cys	261	0.75	83	0.95	28	0.32	28	0.32	122	1.40
Asp	1,503	0.91	211	0.51	443	1.07	305	0.74	544	1.31
Glu	1,626	0.88	374	0.81	703	1.52	98	0.21	451	0.97
Phe	829	0.73	324	1.14	166	0.59	99	0.35	240	0.85
Gly	6,334	3.03	509	0.97	131	0.25	5,069	9.67	625	1.19
His	552	0.83	151	0.91	118	0.71	105	0.63	178	1.07
Ile	862	0.54	363	0.91	244	0.61	2	0.01	253	0.63
Lys	1,766	1.15	521	1.36	641	1.67	99	0.26	505	1.31
Leu	1,539	0.60	759	1.18	304	0.47	61	0.09	415	0.64
Met	363	0.76	116	0.97	75	0.63	26	0.22	146	1.22
Asn	1,124	0.97	224	0.77	181	0.62	504	1.74	215	0.74
Pro	2,083	1.61	620	1.91	1,463	4.51	0	0.00	0	0.00
Gln	981	0.96	268	1.05	266	1.04	77	0.30	370	1.45
Arg	1,171	0.82	381	1.07	353	0.99	108	0.30	329	0.92
Ser	1,327	0.82	294	0.72	335	0.82	109	0.27	589	1.45
Thr	1,144	0.76	389	1.03	238	0.63	35	0.09	482	1.28
Val	1,310	0.66	485	0.97	366	0.73	4	0.01	455	0.91
Trp	294	0.75	106	1.07	57	0.58	25	0.25	106	1.07
Tyr	773	0.79	263	1.08	169	0.69	96	0.39	245	1.00
Other	94	0.68	31	0.90	7	0.20	7	0.20	49	1.42

Table A3.2 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of normal 4 turns. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: n4-3										
Ala	650	0.59	202	0.73	193	0.70	68	0.25	187	0.67
Cys	119	0.74	49	1.21	31	0.77	17	0.42	22	0.54
Asp	991	1.29	166	0.87	580	3.02	130	0.68	115	0.60
Glu	677	0.79	189	0.88	193	0.90	49	0.23	246	1.15
Phe	422	0.81	195	1.49	96	0.73	55	0.42	76	0.58
Gly	3,042	3.14	179	0.74	424	1.75	2,235	9.21	204	0.84
His	396	1.29	117	1.53	144	1.88	42	0.55	93	1.21
Ile	383	0.52	237	1.28	4	0.02	0	0.00	142	0.77
Lys	913	1.28	186	1.05	179	1.01	37	0.21	511	2.87
Leu	548	0.46	240	0.80	80	0.27	27	0.09	201	0.67
Met	148	0.67	36	0.65	41	0.74	22	0.40	49	0.88
Asn	1,282	2.39	124	0.92	756	5.62	271	2.02	131	0.97
Pro	85	0.14	85	0.57	0	0.00	0	0.00	0	0.00
Gln	531	1.12	98	0.83	123	1.04	53	0.45	257	2.17
Arg	640	0.97	166	1.01	122	0.74	46	0.28	306	1.85
Ser	555	0.74	183	0.97	121	0.64	94	0.50	157	0.83
Thr	347	0.50	178	1.02	15	0.09	9	0.05	145	0.83
Val	595	0.65	338	1.47	13	0.06	3	0.01	241	1.04
Trp	122	0.67	58	1.27	20	0.44	15	0.33	29	0.64
Tyr	419	0.93	192	1.70	83	0.73	47	0.42	97	0.86
Other	43	0.67	9	0.56	9	0.56	7	0.44	18	1.13
type: n4-4										
Ala	520	0.70	136	0.73	78	0.42	144	0.77	162	0.87
Cys	125	1.15	48	1.77	26	0.96	34	1.25	17	0.63
Asp	751	1.46	112	0.87	103	0.80	400	3.11	136	1.06
Glu	502	0.87	122	0.85	56	0.39	189	1.31	135	0.94
Phe	289	0.82	138	1.57	15	0.17	46	0.52	90	1.02
Gly	1,999	3.07	136	0.84	1,454	8.93	128	0.79	281	1.73
His	224	1.09	90	1.75	29	0.56	38	0.74	67	1.30
Ile	272	0.55	147	1.18	9	0.07	43	0.35	73	0.59
Lys	420	0.88	99	0.83	71	0.59	114	0.95	136	1.14
Leu	509	0.64	202	1.01	28	0.14	98	0.49	181	0.90
Met	107	0.72	39	1.05	5	0.13	27	0.73	36	0.97
Asn	496	1.38	71	0.79	64	0.71	260	2.88	101	1.12
Pro	139	0.35	59	0.59	1	0.01	79	0.78	0	0.00
Gln	272	0.86	83	1.04	30	0.38	58	0.73	101	1.27
Arg	348	0.79	94	0.85	47	0.42	87	0.79	120	1.08
Ser	577	1.14	127	1.01	95	0.75	198	1.57	157	1.24
Thr	365	0.78	110	0.94	6	0.05	113	0.97	136	1.16
Val	314	0.51	186	1.20	2	0.01	36	0.23	90	0.58
Trp	101	0.83	30	0.98	8	0.26	19	0.62	44	1.44
Tyr	287	0.95	125	1.65	25	0.33	46	0.61	91	1.20
Other	43	1.01	11	1.03	13	1.21	8	0.75	11	1.03

Table A3.2 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of normal 4 turns. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: n4-5										
Ala	64	0.88	16	0.88	29	1.59	0	0.00	19	1.04
Cys	13	1.23	5	1.88	4	1.51	0	0.00	4	1.51
Asp	36	0.72	12	0.95	8	0.63	0	0.00	16	1.27
Glu	24	0.43	7	0.50	11	0.78	0	0.00	6	0.43
Phe	50	1.46	11	1.28	12	1.39	0	0.00	27	3.14
Gly	27	0.42	14	0.88	1	0.06	1	0.06	11	0.69
His	27	1.34	5	0.99	10	1.99	0	0.00	12	2.38
Ile	14	0.29	9	0.74	1	0.08	0	0.00	4	0.33
Lys	51	1.09	9	0.77	33	2.82	0	0.00	9	0.77
Leu	41	0.52	14	0.71	8	0.41	0	0.00	19	0.97
Met	9	0.62	4	1.10	2	0.55	0	0.00	3	0.82
Asn	33	0.94	13	1.47	2	0.23	1	0.11	17	1.92
Pro	260	6.61	18	1.82	33	3.34	209	21.17	0	0.00
Gln	14	0.45	6	0.77	3	0.39	0	0.00	5	0.64
Arg	33	0.76	20	1.85	4	0.37	0	0.00	9	0.83
Ser	31	0.63	11	0.89	8	0.65	0	0.00	12	0.97
Thr	28	0.61	16	1.40	9	0.79	0	0.00	3	0.26
Val	29	0.48	17	1.12	6	0.40	0	0.00	6	0.40
Trp	11	0.92	0	0.00	7	2.33	0	0.00	4	1.33
Tyr	48	1.62	4	0.54	20	2.69	1	0.13	23	3.10
Other	5	1.20	1	0.95	1	0.95	0	0.00	3	2.86
type: n4-6										
Ala	1	0.17	1	0.68	0	0.00	0	0.00	0	0.00
Cys	1	1.17	1	4.70	0	0.00	0	0.00	0	0.00
Asp	6	1.48	1	0.99	1	0.99	3	2.97	1	0.99
Glu	2	0.44	0	0.00	1	0.89	0	0.00	1	0.89
Phe	4	1.45	1	1.45	3	4.35	0	0.00	0	0.00
Gly	10	1.96	6	4.69	0	0.00	3	2.35	1	0.78
His	2	1.24	0	0.00	0	0.00	2	4.95	0	0.00
Ile	1	0.26	1	1.03	0	0.00	0	0.00	0	0.00
Lys	1	0.27	0	0.00	1	1.07	0	0.00	0	0.00
Leu	1	0.16	0	0.00	0	0.00	0	0.00	1	0.64
Met	3	2.57	0	0.00	1	3.42	0	0.00	2	6.85
Asn	2	0.71	1	1.41	1	1.41	0	0.00	0	0.00
Pro	9	2.84	0	0.00	1	1.26	8	10.11	0	0.00
Gln	4	1.60	1	1.60	2	3.20	0	0.00	1	1.60
Arg	2	0.58	0	0.00	0	0.00	0	0.00	2	2.30
Ser	4	1.01	0	0.00	2	2.02	0	0.00	2	2.02
Thr	2	0.55	1	1.09	0	0.00	0	0.00	1	1.09
Val	2	0.41	1	0.82	0	0.00	0	0.00	1	0.82
Trp	4	4.16	0	0.00	3	12.47	1	4.16	0	0.00
Tyr	7	2.94	2	3.36	1	1.68	0	0.00	4	6.71
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

NORMAL 5 TURNS

Table A3.3. Frequencies and conformational parameters of residues within different types and at specific positions of normal 5 turns. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*		*position 4*		*position 5*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: n5-1												
Ala	136,854	1.44	24,381	1.28	27,577	1.45	27,751	1.46	28,150	1.48	28,995	1.52
Cys	12,304	0.89	2,337	0.84	2,268	0.82	2,388	0.86	2,624	0.95	2,687	0.97
Asp	55,645	0.85	14,307	1.09	11,467	0.87	11,393	0.87	9,628	0.73	8,850	0.67
Glu	95,101	1.30	18,763	1.28	20,341	1.38	20,784	1.41	18,597	1.26	16,616	1.13
Phe	44,449	0.99	8,335	0.93	8,728	0.97	8,801	0.98	9,212	1.02	9,373	1.04
Gly	40,888	0.49	10,060	0.60	8,166	0.49	7,501	0.45	6,486	0.39	8,675	0.52
His	22,879	0.87	4,582	0.87	4,340	0.83	4,490	0.85	4,677	0.89	4,790	0.91
Ile	69,408	1.09	12,809	1.01	13,366	1.05	14,101	1.11	14,853	1.17	14,279	1.12
Lys	68,250	1.12	11,960	0.98	13,197	1.08	14,097	1.15	14,339	1.17	14,657	1.20
Leu	134,221	1.31	23,421	1.14	25,661	1.25	26,326	1.29	28,942	1.41	29,871	1.46
Met	24,028	1.26	4,331	1.14	4,666	1.23	4,732	1.24	5,073	1.33	5,226	1.37
Asn	35,291	0.77	8,090	0.88	6,390	0.69	6,599	0.72	6,609	0.72	7,603	0.82
Pro	15,811	0.31	6,766	0.66	5,899	0.57	2,454	0.24	692	0.07	0	0.00
Gln	51,963	1.28	9,728	1.20	10,348	1.27	10,783	1.33	10,768	1.32	10,336	1.27
Arg	68,914	1.22	12,226	1.08	13,354	1.18	13,894	1.23	14,499	1.28	14,941	1.32
Ser	49,929	0.78	12,427	0.96	9,784	0.76	9,332	0.72	8,891	0.69	9,495	0.74
Thr	48,087	0.81	11,358	0.95	9,345	0.78	9,226	0.77	9,662	0.81	8,496	0.71
Val	73,055	0.92	14,069	0.89	14,537	0.92	14,912	0.94	15,335	0.97	14,202	0.90
Trp	15,912	1.02	3,123	1.00	3,315	1.06	3,145	1.00	3,218	1.03	3,111	0.99
Tyr	37,103	0.96	7,061	0.91	7,312	0.94	7,329	0.94	7,701	0.99	7,700	0.99
Other	6,668	1.22	1,218	1.11	1,291	1.18	1,314	1.20	1,396	1.28	1,449	1.32
type: n5-2												
Ala	65	0.53	16	0.66	23	0.95	1	0.04	8	0.33	17	0.70
Cys	16	0.90	8	2.26	0	0.00	0	0.00	5	1.41	3	0.85
Asp	97	1.15	1	0.06	40	2.38	19	1.13	4	0.24	33	1.96
Glu	64	0.68	12	0.64	27	1.44	5	0.27	14	0.74	6	0.32
Phe	37	0.64	14	1.22	6	0.52	4	0.35	6	0.52	7	0.61
Gly	262	2.46	20	0.94	7	0.33	199	9.35	1	0.05	35	1.64
His	26	0.77	3	0.45	4	0.59	0	0.00	11	1.64	8	1.19
Ile	81	1.00	29	1.79	5	0.31	0	0.00	28	1.72	19	1.17
Lys	86	1.10	14	0.90	24	1.54	5	0.32	37	2.37	6	0.38
Leu	120	0.92	43	1.64	16	0.61	2	0.08	14	0.53	45	1.72
Met	21	0.86	7	1.44	4	0.82	1	0.21	2	0.41	7	1.44
Asn	59	1.00	5	0.42	6	0.51	29	2.46	6	0.51	13	1.10
Pro	51	0.77	4	0.30	47	3.57	0	0.00	0	0.00	0	0.00
Gln	50	0.96	16	1.54	10	0.96	3	0.29	16	1.54	5	0.48
Arg	70	0.97	18	1.24	21	1.45	3	0.21	24	1.66	4	0.28
Ser	47	0.57	11	0.67	8	0.49	5	0.30	8	0.49	15	0.91
Thr	68	0.89	12	0.79	6	0.39	1	0.07	30	1.97	19	1.24
Val	124	1.23	31	1.53	16	0.79	0	0.00	53	2.62	24	1.19
Trp	16	0.80	6	1.50	4	1.00	1	0.25	4	1.00	1	0.25
Tyr	50	1.01	8	0.81	9	0.91	5	0.50	12	1.21	16	1.61
Other	5	0.71	5	3.57	0	0.00	0	0.00	0	0.00	0	0.00

Table A3.3 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of normal 5 turns. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: n5-3												
Ala	72	0.77	12	0.64	25	1.33	1	0.05	16	0.85	18	0.96
Cys	13	0.95	4	1.46	3	1.09	0	0.00	3	1.09	3	1.09
Asp	62	0.95	23	1.77	10	0.77	1	0.08	18	1.38	10	0.77
Glu	69	0.95	14	0.96	21	1.44	0	0.00	27	1.86	7	0.48
Phe	46	1.03	10	1.12	19	2.14	0	0.00	12	1.35	5	0.56
Gly	72	0.87	19	1.15	5	0.30	0	0.00	7	0.43	41	2.49
His	38	1.46	9	1.73	5	0.96	2	0.38	10	1.92	12	2.31
Ile	18	0.29	3	0.24	5	0.40	0	0.00	3	0.24	7	0.56
Lys	31	0.51	8	0.66	6	0.50	1	0.08	7	0.58	9	0.75
Leu	57	0.56	18	0.89	9	0.44	0	0.00	16	0.79	14	0.69
Met	12	0.64	2	0.53	4	1.06	0	0.00	1	0.27	5	1.33
Asn	48	1.05	13	1.42	10	1.10	1	0.11	8	0.88	16	1.75
Pro	249	4.88	16	1.57	27	2.65	206	20.20	0	0.00	0	0.00
Gln	34	0.85	9	1.12	7	0.87	0	0.00	15	1.87	3	0.37
Arg	30	0.54	6	0.54	5	0.45	1	0.09	9	0.80	9	0.80
Ser	49	0.77	12	0.94	11	0.86	1	0.08	8	0.63	17	1.33
Thr	47	0.80	14	1.19	4	0.34	3	0.25	17	1.44	9	0.76
Val	45	0.58	12	0.77	9	0.58	2	0.13	6	0.38	16	1.02
Trp	25	1.61	3	0.97	6	1.94	0	0.00	10	3.23	6	1.94
Tyr	74	1.93	12	1.56	26	3.39	0	0.00	24	3.13	12	1.56
Other	4	0.74	0	0.00	2	1.85	0	0.00	2	1.85	0	0.00
type: n5-4												
Ala	84	1.03	8	0.49	6	0.37	19	1.16	18	1.10	33	2.02
Cys	6	0.50	1	0.42	0	0.00	2	0.84	3	1.26	0	0.00
Asp	44	0.78	14	1.24	6	0.53	14	1.24	9	0.80	1	0.09
Glu	60	0.95	9	0.71	3	0.24	20	1.58	15	1.19	13	1.03
Phe	27	0.70	5	0.65	1	0.13	7	0.91	11	1.43	3	0.39
Gly	319	4.46	56	3.92	156	10.92	6	0.42	26	1.82	75	5.25
His	16	0.71	7	1.55	1	0.22	2	0.44	3	0.66	3	0.66
Ile	38	0.70	6	0.55	0	0.00	27	2.48	5	0.46	0	0.00
Lys	45	0.86	12	1.15	2	0.19	11	1.05	8	0.76	12	1.15
Leu	64	0.73	8	0.46	0	0.00	23	1.31	18	1.02	15	0.85
Met	12	0.73	2	0.61	0	0.00	6	1.84	4	1.22	0	0.00
Asn	16	0.40	2	0.25	2	0.25	7	0.88	4	0.51	1	0.13
Pro	17	0.38	6	0.68	0	0.00	6	0.68	5	0.57	0	0.00
Gln	26	0.75	8	1.15	1	0.14	4	0.57	5	0.72	8	1.15
Arg	39	0.80	9	0.93	0	0.00	6	0.62	9	0.93	15	1.54
Ser	47	0.85	18	1.63	10	0.90	7	0.63	8	0.72	4	0.36
Thr	35	0.68	10	0.98	1	0.10	5	0.49	17	1.66	2	0.20
Val	23	0.34	3	0.22	0	0.00	11	0.81	6	0.44	3	0.22
Trp	12	0.89	2	0.74	0	0.00	4	1.49	5	1.86	1	0.37
Tyr	16	0.48	3	0.45	1	0.15	1	0.15	10	1.50	1	0.15
Other	4	0.85	1	1.06	0	0.00	2	2.13	1	1.06	0	0.00

Table A3.3 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of normal 5 turns. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: n5-5												
Ala	8	0.33	1	0.20	3	0.61	0	0.00	2	0.41	2	0.41
Cys	1	0.28	0	0.00	1	1.40	0	0.00	0	0.00	0	0.00
Asp	13	0.77	1	0.30	7	2.07	4	1.18	0	0.00	1	0.30
Glu	15	0.79	1	0.26	5	1.32	4	1.06	4	1.06	1	0.26
Phe	14	1.21	6	2.59	2	0.86	2	0.86	0	0.00	4	1.73
Gly	47	2.19	8	1.87	7	1.63	28	6.53	1	0.23	3	0.70
His	9	1.33	3	2.22	2	1.48	0	0.00	1	0.74	3	2.22
Ile	17	1.04	5	1.53	0	0.00	0	0.00	4	1.22	8	2.45
Lys	16	1.02	2	0.64	2	0.64	2	0.64	9	2.86	1	0.32
Leu	20	0.76	11	2.09	1	0.19	2	0.38	2	0.38	4	0.76
Met	7	1.43	0	0.00	2	2.04	1	1.02	1	1.02	3	3.06
Asn	24	2.02	0	0.00	15	6.32	7	2.95	1	0.42	1	0.42
Pro	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Gln	16	1.53	1	0.48	3	1.43	3	1.43	7	3.34	2	0.96
Arg	16	1.10	2	0.69	3	1.03	0	0.00	10	3.43	1	0.34
Ser	9	0.54	1	0.30	2	0.60	1	0.30	2	0.60	3	0.90
Thr	6	0.39	1	0.33	0	0.00	0	0.00	5	1.63	0	0.00
Val	24	1.18	8	1.96	0	0.00	0	0.00	6	1.47	10	2.45
Trp	2	0.50	1	1.24	1	1.24	0	0.00	0	0.00	0	0.00
Tyr	20	2.00	5	2.50	1	0.50	2	1.00	2	1.00	10	5.01
Other	1	0.71	0	0.00	0	0.00	1	3.55	0	0.00	0	0.00
type: n5-6												
Ala	10	0.45	4	0.90	5	1.12	1	0.22	0	0.00	0	0.00
Cys	3	0.92	1	1.54	2	3.07	0	0.00	0	0.00	0	0.00
Asp	20	1.29	2	0.65	8	2.59	3	0.97	5	1.62	2	0.65
Glu	9	0.52	4	1.16	1	0.29	3	0.87	0	0.00	1	0.29
Phe	7	0.66	2	0.95	3	1.42	2	0.95	0	0.00	0	0.00
Gly	70	3.58	9	2.30	3	0.77	28	7.16	18	4.60	12	3.07
His	2	0.32	0	0.00	0	0.00	1	0.81	0	0.00	1	0.81
Ile	7	0.47	2	0.67	2	0.67	0	0.00	0	0.00	3	1.01
Lys	11	0.77	1	0.35	3	1.05	0	0.00	1	0.35	6	2.09
Leu	24	1.00	16	3.33	0	0.00	2	0.42	4	0.83	2	0.42
Met	4	0.90	0	0.00	2	2.24	0	0.00	0	0.00	2	2.24
Asn	28	2.58	2	0.92	1	0.46	8	3.69	15	6.92	2	0.92
Pro	10	0.83	0	0.00	10	4.13	0	0.00	0	0.00	0	0.00
Gln	11	1.15	1	0.52	2	1.05	0	0.00	3	1.57	5	2.62
Arg	10	0.75	1	0.38	0	0.00	0	0.00	1	0.38	8	3.01
Ser	8	0.53	2	0.66	4	1.32	0	0.00	1	0.33	1	0.33
Thr	11	0.78	3	1.07	2	0.71	0	0.00	1	0.36	5	1.78
Val	1	0.05	0	0.00	0	0.00	0	0.00	0	0.00	1	0.27
Trp	4	1.09	2	2.72	1	1.36	0	0.00	1	1.36	0	0.00
Tyr	10	1.10	0	0.00	3	1.65	4	2.19	2	1.10	1	0.55
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

NORMAL 6 TURNS

Table A3.4. Frequencies and conformational parameters of residues within different types and at specific positions of normal 6 turns. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*		*position 4*		*position 5*		*position 6*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: n6-1														
Ala	6,066	1.40	1,860	2.57	884	1.22	1,279	1.77	1,161	1.60	121	0.17	761	1.05
Cys	625	0.99	285	2.70	42	0.40	40	0.38	100	0.95	21	0.20	137	1.30
Asp	2,201	0.73	381	0.76	368	0.73	591	1.18	237	0.47	382	0.76	242	0.48
Glu	3,387	1.01	202	0.36	906	1.62	1,191	2.13	525	0.94	244	0.44	319	0.57
Phe	1,578	0.77	402	1.17	195	0.57	124	0.36	358	1.05	30	0.09	469	1.37
Gly	6,891	1.81	303	0.48	170	0.27	227	0.36	182	0.29	5,354	8.45	655	1.03
His	1,139	0.95	118	0.59	251	1.25	139	0.69	315	1.57	182	0.91	134	0.67
Ile	2,253	0.78	347	0.72	532	1.10	211	0.44	238	0.49	6	0.01	919	1.90
Lys	3,716	1.33	194	0.42	975	2.10	966	2.08	583	1.25	508	1.09	490	1.05
Leu	6,107	1.31	1,958	2.51	833	1.07	623	0.80	1,397	1.79	42	0.05	1,254	1.61
Met	996	1.15	283	1.95	137	0.95	133	0.92	230	1.59	23	0.16	190	1.31
Asn	2,286	1.09	214	0.61	210	0.60	324	0.92	580	1.65	770	2.19	188	0.54
Pro	354	0.15	111	0.28	172	0.44	66	0.17	5	0.01	0	0.00	0	0.00
Gln	2,126	1.15	156	0.50	477	1.54	553	1.79	470	1.52	240	0.78	230	0.74
Arg	3,027	1.17	277	0.64	742	1.72	730	1.69	613	1.42	294	0.68	371	0.86
Ser	1,957	0.66	362	0.74	321	0.65	537	1.09	352	0.72	146	0.30	239	0.49
Thr	1,326	0.49	148	0.33	320	0.70	272	0.60	314	0.69	14	0.03	258	0.57
Val	2,365	0.66	331	0.55	568	0.94	225	0.37	259	0.43	9	0.01	973	1.61
Trp	475	0.66	119	1.00	89	0.75	47	0.39	91	0.76	5	0.04	124	1.04
Tyr	1,420	0.80	291	0.98	205	0.69	127	0.43	360	1.22	25	0.08	412	1.39
Other	285	1.14	88	2.11	33	0.79	25	0.60	60	1.44	14	0.34	65	1.56
type: n6-2														
Ala	863	0.78	155	0.84	195	1.05	145	0.78	91	0.49	123	0.66	154	0.83
Cys	181	1.12	20	0.74	33	1.22	29	1.07	34	1.26	34	1.26	31	1.15
Asp	664	0.86	81	0.63	115	0.90	123	0.96	151	1.18	96	0.75	98	0.76
Glu	949	1.10	105	0.73	189	1.32	219	1.53	182	1.27	142	0.99	112	0.78
Phe	815	1.55	218	2.49	118	1.35	100	1.14	117	1.34	124	1.42	138	1.58
Gly	773	0.79	59	0.36	67	0.41	83	0.51	123	0.76	106	0.65	335	2.07
His	347	1.13	54	1.05	46	0.90	51	1.00	92	1.80	62	1.21	42	0.82
Ile	919	1.24	209	1.69	160	1.29	102	0.82	103	0.83	217	1.75	128	1.03
Lys	683	0.96	64	0.54	98	0.82	170	1.43	124	1.04	108	0.91	119	1.00
Leu	1,438	1.20	267	1.34	259	1.30	160	0.80	167	0.84	264	1.32	321	1.61
Met	295	1.33	65	1.75	75	2.02	51	1.38	22	0.59	34	0.92	48	1.29
Asn	577	1.07	64	0.71	88	0.98	147	1.64	143	1.59	63	0.70	72	0.80
Pro	87	0.14	30	0.30	38	0.38	12	0.12	0	0.00	7	0.07	0	0.00
Gln	404	0.85	64	0.81	76	0.96	86	1.09	69	0.87	54	0.68	55	0.69
Arg	550	0.83	84	0.76	108	0.98	126	1.14	100	0.91	64	0.58	68	0.62
Ser	533	0.71	90	0.72	88	0.70	114	0.91	104	0.83	70	0.56	67	0.53
Thr	661	0.95	85	0.73	92	0.79	140	1.20	181	1.56	99	0.85	64	0.55
Val	1,077	1.17	169	1.10	143	0.93	156	1.01	178	1.15	287	1.86	144	0.93
Trp	262	1.43	66	2.16	46	1.51	26	0.85	34	1.11	45	1.47	45	1.47
Tyr	791	1.75	192	2.54	108	1.43	106	1.40	131	1.73	149	1.97	105	1.39
Other	73	1.14	16	1.50	15	1.41	11	1.03	11	1.03	9	0.84	11	1.03

Table A3.4 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of normal 6 turns. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: n6-3														
Ala	41	0.99	13	1.89	7	1.02	3	0.44	3	0.44	8	1.16	7	1.02
Cys	5	0.83	1	1.00	0	0.00	1	1.00	1	1.00	0	0.00	2	2.00
Asp	21	0.74	3	0.63	10	2.10	4	0.84	0	0.00	0	0.00	4	0.84
Glu	14	0.44	2	0.38	4	0.75	3	0.56	1	0.19	2	0.38	2	0.38
Phe	16	0.82	3	0.92	3	0.92	3	0.92	3	0.92	1	0.31	3	0.92
Gly	101	2.80	1	0.17	11	1.83	20	3.32	55	9.14	5	0.83	9	1.50
His	7	0.61	0	0.00	2	1.05	1	0.53	0	0.00	4	2.10	0	0.00
Ile	17	0.62	5	1.09	0	0.00	3	0.65	0	0.00	3	0.65	6	1.31
Lys	19	0.72	1	0.23	2	0.45	1	0.23	1	0.23	8	1.81	6	1.36
Leu	33	0.74	11	1.49	6	0.81	2	0.27	1	0.14	6	0.81	7	0.95
Met	5	0.61	0	0.00	1	0.73	1	0.73	0	0.00	2	1.45	1	0.73
Asn	47	2.35	1	0.30	13	3.90	24	7.20	4	1.20	2	0.60	3	0.90
Pro	3	0.13	2	0.54	0	0.00	0	0.00	0	0.00	1	0.27	0	0.00
Gln	18	1.02	2	0.68	5	1.70	1	0.34	1	0.34	6	2.04	3	1.02
Arg	18	0.73	2	0.49	1	0.24	1	0.24	1	0.24	9	2.20	4	0.98
Ser	35	1.25	10	2.15	7	1.50	4	0.86	3	0.64	6	1.29	5	1.07
Thr	32	1.24	11	2.55	2	0.46	1	0.23	0	0.00	10	2.32	8	1.85
Val	16	0.47	2	0.35	2	0.35	1	0.17	0	0.00	4	0.70	7	1.22
Trp	10	1.47	2	1.77	1	0.88	4	3.53	2	1.77	1	0.88	0	0.00
Tyr	19	1.13	7	2.50	3	1.07	2	0.71	3	1.07	1	0.36	3	1.07
Other	3	1.26	1	2.53	0	0.00	0	0.00	1	2.53	1	2.53	0	0.00
type: n6-4														
Ala	17	0.49	2	0.35	4	0.69	1	0.17	3	0.52	3	0.52	4	0.69
Cys	4	0.79	3	3.58	0	0.00	0	0.00	0	0.00	0	0.00	1	1.19
Asp	48	2.01	2	0.50	29	7.28	7	1.76	1	0.25	7	1.76	2	0.50
Glu	8	0.30	0	0.00	2	0.45	2	0.45	2	0.45	2	0.45	0	0.00
Phe	18	1.10	5	1.84	5	1.84	0	0.00	0	0.00	3	1.10	5	1.84
Gly	61	2.02	4	0.79	1	0.20	46	9.13	8	1.59	0	0.00	2	0.40
His	3	0.31	0	0.00	1	0.63	0	0.00	1	0.63	1	0.63	0	0.00
Ile	26	1.13	8	2.08	1	0.26	0	0.00	1	0.26	10	2.60	6	1.56
Lys	6	0.27	1	0.27	2	0.54	0	0.00	2	0.54	1	0.27	0	0.00
Leu	53	1.43	14	2.26	2	0.32	0	0.00	3	0.48	22	3.55	12	1.94
Met	7	1.01	3	2.61	1	0.87	0	0.00	0	0.00	2	1.74	1	0.87
Asn	13	0.78	2	0.72	0	0.00	4	1.43	1	0.36	2	0.72	4	1.43
Pro	1	0.05	0	0.00	1	0.32	0	0.00	0	0.00	0	0.00	0	0.00
Gln	10	0.68	3	1.22	2	0.81	0	0.00	1	0.41	2	0.81	2	0.81
Arg	12	0.58	5	1.46	2	0.58	0	0.00	4	1.17	1	0.29	0	0.00
Ser	11	0.47	0	0.00	3	0.77	2	0.51	1	0.26	4	1.02	1	0.26
Thr	40	1.84	4	1.11	2	0.55	0	0.00	25	6.92	2	0.55	7	1.94
Val	38	1.32	6	1.25	4	0.84	0	0.00	13	2.71	4	0.84	11	2.30
Trp	12	2.11	2	2.11	2	2.11	0	0.00	0	0.00	1	1.05	7	7.38
Tyr	11	0.78	0	0.00	3	1.28	5	2.13	1	0.43	0	0.00	2	0.85
Other	3	1.51	3	9.06	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

OPEN 4 TURNS

Table A3.5. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 4 turn structures. Turn types o4-23 and o4-33 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		ψ_{i+1}		φ_{i+1}		ω_{i+1}		ψ_{i+2}		φ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o4-23	87	178.0	6.2	-162.7	18.4	-177.6	10.7	177.7	6.2	-69.6	14.6	97.6	22.1
o4-23	*53*	*177.2*	*6.9*	*-159.4*	*16.9*	*-177.2*	*9.0*	*177.4*	*6.0*	*-69.9*	*13.6*	*97.4*	*22.5*
o4-24	67	178.2	5.7	93.6	17.2	146.6	28.0	177.5	4.7	63.7	13.7	27.2	16.4
o4-24	*56*	*177.6*	*8.7*	*94.3*	*17.3*	*146.6*	*28.7*	*176.6*	*11.6*	*64.2*	*13.8*	*26.0*	*15.4*
o4-25	71	179.2	5.0	-83.9	17.0	-21.1	19.3	179.7	5.1	-127.7	19.2	134.0	13.4
o4-25	*48*	*179.4*	*5.8*	*-85.2*	*17.5*	*-23.6*	*19.5*	*178.5*	*4.1*	*-128.3*	*20.5*	*132.5*	*13.2*
o4-26	68	174.7	5.7	-74.0	12.3	153.2	18.4	177.0	5.9	-82.4	26.0	128.9	17.2
o4-26	*42*	*173.9*	*5.8*	*-73.8*	*10.9*	*153.2*	*20.7*	*178.0*	*5.3*	*-86.5*	*27.4*	*124.8*	*18.1*
o4-27	54	-177.9	5.9	-85.6	5.3	64.4	5.6	-175.4	5.0	-126.6	13.2	16.5	10.8
o4-27	*48*	*-178.2*	*5.7*	*-85.8*	*7.9*	*65.2*	*6.8*	*-176.3*	*5.0*	*-126.2*	*13.9*	*15.4*	*11.1*
o4-28	62	179.7	4.6	74.9	11.7	11.3	14.3	-177.7	4.7	-120.3	7.9	-49.5	8.7
o4-28	*43*	*179.0*	*5.0*	*74.3*	*11.7*	*12.1*	*15.8*	*-177.6*	*4.4*	*-120.7*	*9.0*	*-49.5*	*10.2*
o4-29	45	179.1	5.2	-96.6	19.6	127.9	8.9	176.4	7.3	75.4	7.5	-60.7	13.4
o4-29	*31*	*179.3*	*4.8*	*-95.2*	*16.4*	*125.7*	*10.1*	*176.2*	*6.7*	*73.7*	*6.9*	*-61.8*	*17.0*
o4-30	48	179.9	4.2	-135.4	14.2	137.0	10.2	2.4	4.3	-79.8	7.1	-11.1	9.3
o4-30	*37*	*178.7*	*3.9*	*-135.0*	*14.7*	*134.8*	*11.0*	*3.4*	*4.6*	*-79.4*	*6.9*	*-12.2*	*8.7*
o4-31	52	3.3	5.1	-91.0	5.4	10.9	11.6	-177.0	5.5	-70.8	16.1	-22.9	21.1
o4-31	*35*	*3.4*	*4.4*	*-92.9*	*5.6*	*11.6*	*12.0*	*-177.2*	*6.8*	*-72.5*	*15.8*	*-23.5*	*19.3*
o4-32	40	179.9	3.7	88.6	10.2	1.7	10.4	179.7	4.0	101.1	14.1	150.0	15.9
o4-32	*27*	*-179.8*	*4.6*	*90.7*	*9.4*	*0.5*	*9.5*	*179.4*	*3.8*	*102.3*	*12.6*	*147.6*	*17.3*
o4-33	37	177.8	4.0	-60.0	6.3	146.9	6.5	2.3	3.4	-86.7	6.5	-4.1	10.7
o4-33	*21*	*178.3*	*4.5*	*-60.4*	*6.9*	*148.7*	*6.8*	*1.1*	*3.7*	*-86.4*	*8.2*	*-4.7*	*12.4*

type	|turns|	ω_{i+2}		Cα-Cα[Å]	
		∅	±	∅	±
o4-23	87	-175.0	8.3	7.8	0.2
o4-23	*53*	*-174.0*	*5.8*	*7.8*	*0.2*
o4-24	67	-179.4	4.6	7.2	0.6
o4-24	*56*	*179.7*	*9.7*	*7.2*	*0.6*
o4-25	71	-1.6	6.0	7.5	0.4
o4-25	*48*	*-1.3*	*6.9*	*7.4*	*0.5*
o4-26	68	-2.5	6.2	7.7	0.3
o4-26	*42*	*-3.3*	*8.6*	*7.7*	*0.3*
o4-27	54	-178.5	4.8	7.8	0.2
o4-27	*48*	*-178.8*	*4.7*	*7.8*	*0.2*
o4-28	62	-178.2	6.1	6.9	0.5
o4-28	*43*	*-177.9*	*6.9*	*6.9*	*0.4*
o4-29	45	-179.3	6.8	6.3	0.6
o4-29	*31*	*179.7*	*7.0*	*6.4*	*0.5*
o4-30	48	-179.8	4.7	5.5	0.7
o4-30	*37*	*-179.8*	*5.6*	*5.4*	*0.7*
o4-31	52	178.4	5.2	5.4	0.5
o4-31	*35*	*177.8*	*5.5*	*5.4*	*0.5*
o4-32	40	-179.1	4.7	7.5	0.5
o4-32	*27*	*-177.8*	*4.6*	*7.5*	*0.3*
o4-33	37	179.9	4.5	5.9	0.4
o4-33	*21*	*179.9*	*3.2*	*6.0*	*0.5*

Table A3.6. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*		*position 4*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-1										
Ala	11,793	0.87	3,680	1.08	2,942	0.86	2,708	0.80	2,463	0.72
Cys	2,063	1.04	632	1.27	347	0.70	650	1.31	434	0.87
Asp	11,998	1.27	2,078	0.88	3,996	1.70	3,367	1.43	2,557	1.09
Glu	9,424	0.90	2,182	0.83	3,053	1.16	2,147	0.82	2,042	0.78
Phe	6,374	0.99	1,528	0.95	1,059	0.66	2,428	1.51	1,359	0.84
Gly	8,499	0.71	3,516	1.18	1,206	0.40	1,719	0.58	2,058	0.69
His	3,870	1.03	791	0.84	983	1.04	1,415	1.50	681	0.72
Ile	8,630	0.95	2,241	0.99	1,551	0.68	2,491	1.10	2,347	1.03
Lys	9,277	1.06	1,928	0.88	3,227	1.48	1,880	0.86	2,242	1.03
Leu	11,660	0.80	2,982	0.81	3,114	0.85	2,962	0.81	2,602	0.71
Met	2,049	0.75	510	0.75	479	0.70	579	0.85	481	0.71
Asn	7,495	1.14	1,433	0.87	2,122	1.29	2,541	1.54	1,399	0.85
Pro	12,052	1.63	3,485	1.89	1,777	0.96	48	0.03	6,742	3.65
Gln	5,363	0.92	1,210	0.83	1,753	1.21	1,312	0.90	1,088	0.75
Arg	7,740	0.96	1,685	0.83	2,432	1.20	1,930	0.95	1,693	0.84
Ser	11,450	1.24	3,040	1.32	3,030	1.31	3,051	1.32	2,329	1.01
Thr	9,575	1.12	2,014	0.94	2,858	1.34	2,400	1.12	2,303	1.08
Val	11,214	0.99	2,972	1.05	2,036	0.72	3,110	1.10	3,096	1.09
Trp	1,974	0.88	414	0.74	436	0.78	615	1.10	509	0.91
Tyr	5,339	0.96	1,133	0.82	1,075	0.77	2,093	1.51	1,038	0.75
Other	645	0.82	167	0.85	145	0.74	175	0.89	158	0.81
type: o4-2										
Ala	8,339	0.98	2,117	0.99	1,973	0.93	2,197	1.03	2,052	0.96
Cys	1,266	1.02	349	1.12	313	1.01	318	1.02	286	0.92
Asp	7,599	1.29	2,545	1.73	1,685	1.14	1,707	1.16	1,662	1.13
Glu	6,819	1.04	1,922	1.17	1,832	1.11	1,629	0.99	1,436	0.87
Phe	4,121	1.02	706	0.70	1,237	1.23	1,310	1.30	868	0.86
Gly	6,870	0.92	1,925	1.03	1,011	0.54	1,230	0.66	2,704	1.45
His	2,840	1.21	648	1.10	723	1.23	774	1.31	695	1.18
Ile	3,787	0.67	730	0.51	1,083	0.76	1,012	0.71	962	0.68
Lys	5,530	1.01	1,271	0.93	1,401	1.03	1,434	1.05	1,424	1.04
Leu	8,584	0.94	1,720	0.75	2,872	1.25	2,493	1.09	1,499	0.65
Met	1,612	0.95	357	0.84	497	1.17	420	0.99	338	0.79
Asn	5,209	1.26	1,400	1.36	1,224	1.19	1,438	1.39	1,147	1.11
Pro	4,617	1.00	1,361	1.18	901	0.78	162	0.14	2,193	1.90
Gln	3,848	1.06	907	1.00	1,049	1.15	969	1.07	923	1.01
Arg	5,117	1.01	1,269	1.00	1,219	0.96	1,429	1.13	1,200	0.95
Ser	7,550	1.31	2,353	1.63	1,562	1.08	1,934	1.34	1,701	1.18
Thr	5,407	1.01	1,248	0.93	1,470	1.10	1,502	1.12	1,187	0.89
Val	4,345	0.61	833	0.47	1,160	0.65	1,148	0.65	1,204	0.68
Trp	1,612	1.15	327	0.93	436	1.24	459	1.31	390	1.11
Tyr	3,634	1.05	688	0.79	1,005	1.16	1,113	1.28	828	0.95
Other	418	0.85	105	0.86	128	1.04	103	0.84	82	0.67

Table A3.6 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-3										
Ala	1,610	0.67	428	0.71	438	0.73	279	0.46	465	0.77
Cys	425	1.22	117	1.34	77	0.88	68	0.78	163	1.86
Asp	1,816	1.09	207	0.50	386	0.93	1,029	2.48	194	0.47
Glu	1,478	0.80	378	0.81	558	1.20	346	0.75	196	0.42
Phe	1,157	1.02	438	1.54	300	1.06	176	0.62	243	0.86
Gly	4,394	2.09	491	0.93	202	0.38	1,937	3.69	1,764	3.36
His	742	1.12	139	0.84	211	1.27	271	1.63	121	0.73
Ile	1,401	0.87	615	1.53	438	1.09	5	0.01	343	0.86
Lys	1,248	0.81	289	0.75	495	1.28	227	0.59	237	0.61
Leu	1,798	0.70	781	1.21	433	0.67	149	0.23	435	0.67
Met	392	0.82	144	1.20	95	0.79	57	0.47	96	0.80
Asn	2,295	1.97	200	0.69	255	0.88	1,499	5.15	341	1.17
Pro	1,281	0.98	409	1.26	540	1.66	0	0.00	332	1.02
Gln	892	0.87	210	0.82	276	1.08	176	0.69	230	0.90
Arg	1,187	0.83	311	0.87	431	1.21	227	0.64	218	0.61
Ser	1,418	0.87	295	0.72	450	1.10	248	0.61	425	1.04
Thr	1,226	0.81	411	1.09	419	1.11	30	0.08	366	0.97
Val	1,887	0.95	712	1.43	576	1.15	38	0.08	561	1.12
Trp	306	0.77	98	0.99	94	0.95	47	0.48	67	0.68
Tyr	919	0.94	281	1.15	304	1.24	168	0.69	166	0.68
Other	84	0.61	35	1.01	11	0.32	12	0.35	26	0.75
type: o4-4										
Ala	2,060	0.86	550	0.92	183	0.31	737	1.23	590	0.99
Cys	293	0.84	146	1.67	66	0.76	43	0.49	38	0.44
Asp	3,014	1.82	145	0.35	1,238	2.99	602	1.45	1,029	2.48
Glu	2,127	1.15	236	0.51	109	0.24	835	1.80	947	2.04
Phe	893	0.79	486	1.72	62	0.22	151	0.53	194	0.69
Gly	2,196	1.05	466	0.89	1,122	2.14	231	0.44	377	0.72
His	568	0.86	124	0.75	114	0.69	145	0.88	185	1.12
Ile	761	0.48	362	0.91	15	0.04	229	0.57	155	0.39
Lys	1,302	0.85	319	0.83	111	0.29	444	1.15	428	1.11
Leu	1,996	0.77	1,177	1.82	136	0.21	370	0.57	313	0.49
Met	352	0.73	183	1.53	34	0.28	64	0.53	71	0.59
Asn	1,593	1.37	151	0.52	668	2.30	227	0.78	547	1.88
Pro	1,792	1.38	470	1.45	292	0.90	993	3.06	37	0.11
Gln	774	0.76	175	0.68	50	0.20	216	0.84	333	1.30
Arg	1,057	0.74	313	0.88	99	0.28	315	0.88	330	0.93
Ser	2,626	1.62	298	0.73	1,242	3.06	569	1.40	517	1.27
Thr	2,329	1.55	372	0.99	1,322	3.52	259	0.69	376	1.00
Val	1,000	0.50	514	1.03	30	0.06	270	0.54	186	0.37
Trp	343	0.87	126	1.28	19	0.19	98	0.99	100	1.01
Tyr	715	0.73	312	1.28	52	0.21	149	0.61	202	0.83
Other	93	0.68	46	1.33	7	0.20	24	0.70	16	0.46

Table A3.6 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-5										
Ala	912	0.78	299	1.02	58	0.20	282	0.97	273	0.94
Cys	166	0.98	45	1.06	17	0.40	35	0.82	69	1.62
Asp	718	0.89	271	1.34	99	0.49	181	0.90	167	0.83
Glu	753	0.83	240	1.06	61	0.27	215	0.95	237	1.05
Phe	434	0.79	78	0.57	53	0.38	129	0.93	174	1.26
Gly	3,194	3.13	174	0.68	2,439	9.54	88	0.34	493	1.93
His	290	0.90	91	1.13	50	0.62	80	0.99	69	0.85
Ile	494	0.63	93	0.48	0	0.00	164	0.84	237	1.22
Lys	736	0.98	233	1.24	94	0.50	320	1.71	89	0.47
Leu	942	0.75	209	0.66	40	0.13	327	1.04	366	1.16
Met	194	0.83	50	0.86	14	0.24	81	1.39	49	0.84
Asn	661	1.17	189	1.33	197	1.39	158	1.12	117	0.83
Pro	430	0.68	340	2.15	0	0.00	2	0.01	88	0.56
Gln	462	0.93	154	1.23	34	0.27	187	1.50	87	0.70
Arg	629	0.91	185	1.06	94	0.54	244	1.40	106	0.61
Ser	798	1.01	319	1.61	51	0.26	253	1.28	175	0.88
Thr	657	0.90	207	1.13	21	0.11	258	1.41	171	0.93
Val	543	0.56	86	0.35	5	0.02	208	0.86	244	1.00
Trp	150	0.78	32	0.67	15	0.31	52	1.08	51	1.06
Tyr	382	0.80	94	0.79	52	0.44	112	0.94	124	1.04
Other	47	0.70	9	0.54	4	0.24	22	1.31	12	0.71
type: o4-6										
Ala	659	0.64	228	0.88	167	0.64	69	0.27	195	0.75
Cys	170	1.13	56	1.48	53	1.40	28	0.74	33	0.87
Asp	1,091	1.52	260	1.45	586	3.27	99	0.55	146	0.81
Glu	563	0.70	216	1.08	121	0.60	43	0.21	183	0.91
Phe	365	0.75	89	0.73	117	0.96	48	0.39	111	0.91
Gly	2,923	3.23	431	1.90	77	0.34	2,088	9.21	327	1.44
His	243	0.85	60	0.84	57	0.80	50	0.70	76	1.06
Ile	301	0.44	77	0.45	54	0.31	5	0.03	165	0.95
Lys	628	0.95	161	0.97	118	0.71	82	0.49	267	1.61
Leu	670	0.60	178	0.64	190	0.68	55	0.20	247	0.89
Met	128	0.62	28	0.54	39	0.75	25	0.48	36	0.69
Asn	758	1.51	172	1.37	253	2.01	187	1.49	146	1.16
Pro	230	0.41	184	1.31	24	0.17	0	0.00	22	0.16
Gln	367	0.83	101	0.91	83	0.75	54	0.49	129	1.17
Arg	463	0.75	154	1.00	108	0.70	76	0.49	125	0.81
Ser	905	1.29	207	1.18	393	2.24	58	0.33	247	1.41
Thr	759	1.17	152	0.93	389	2.39	9	0.06	209	1.29
Val	341	0.40	94	0.44	62	0.29	4	0.02	181	0.84
Trp	137	0.80	56	1.31	35	0.82	5	0.12	41	0.96
Tyr	320	0.76	100	0.95	80	0.76	24	0.23	116	1.10
Other	35	0.59	10	0.67	8	0.54	5	0.34	12	0.81

Table A3.6 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-7										
Ala	312	0.63	114	0.92	88	0.71	34	0.27	76	0.61
Cys	53	0.73	14	0.78	21	1.16	9	0.50	9	0.50
Asp	459	1.34	21	0.25	93	1.09	69	0.81	276	3.22
Glu	333	0.87	52	0.54	123	1.28	39	0.41	119	1.24
Phe	166	0.71	59	1.01	74	1.26	4	0.07	29	0.50
Gly	1,322	3.05	188	1.73	22	0.20	1,031	9.51	81	0.75
His	134	0.98	34	0.99	60	1.75	14	0.41	26	0.76
Ile	252	0.76	90	1.09	126	1.52	4	0.05	32	0.39
Lys	317	1.00	121	1.52	89	1.12	33	0.42	74	0.93
Leu	269	0.50	86	0.65	110	0.83	11	0.08	62	0.47
Met	65	0.66	24	0.97	18	0.73	4	0.16	19	0.77
Asn	314	1.31	23	0.38	51	0.85	42	0.70	198	3.30
Pro	165	0.61	65	0.97	52	0.77	0	0.00	48	0.72
Gln	176	0.83	65	1.23	56	1.06	14	0.26	41	0.77
Arg	264	0.90	113	1.53	64	0.87	29	0.39	58	0.79
Ser	356	1.06	76	0.91	83	0.99	72	0.86	125	1.49
Thr	252	0.81	96	1.24	57	0.73	3	0.04	96	1.24
Val	308	0.75	119	1.16	159	1.54	3	0.03	27	0.26
Trp	56	0.69	24	1.18	19	0.93	2	0.10	11	0.54
Tyr	168	0.83	52	1.03	69	1.37	14	0.28	33	0.65
Other	23	0.81	5	0.70	7	0.98	10	1.40	1	0.14
type: o4-8										
Ala	345	0.75	106	0.93	9	0.08	108	0.94	122	1.07
Cys	56	0.84	27	1.62	1	0.06	8	0.48	20	1.20
Asp	335	1.06	56	0.71	15	0.19	115	1.45	149	1.88
Glu	315	0.89	66	0.75	4	0.05	124	1.40	121	1.37
Phe	156	0.72	61	1.13	1	0.02	45	0.83	49	0.91
Gly	1,612	4.02	177	1.77	1,223	12.20	36	0.36	176	1.76
His	108	0.85	43	1.36	1	0.03	23	0.73	41	1.29
Ile	181	0.59	63	0.82	0	0.00	76	0.99	42	0.55
Lys	191	0.65	55	0.75	4	0.05	68	0.93	64	0.87
Leu	267	0.54	113	0.92	1	0.01	94	0.76	59	0.48
Met	64	0.70	23	1.00	1	0.04	23	1.00	17	0.74
Asn	175	0.79	47	0.85	9	0.16	53	0.95	66	1.19
Pro	232	0.94	41	0.66	0	0.00	171	2.75	20	0.32
Gln	145	0.74	47	0.96	0	0.00	46	0.94	52	1.06
Arg	178	0.65	64	0.94	4	0.06	59	0.87	51	0.75
Ser	313	1.01	92	1.18	34	0.44	105	1.35	82	1.06
Thr	241	0.84	73	1.02	21	0.29	66	0.92	81	1.13
Val	182	0.48	69	0.72	0	0.00	51	0.54	62	0.65
Trp	67	0.89	32	1.70	1	0.05	22	1.17	12	0.64
Tyr	148	0.79	69	1.48	1	0.02	35	0.75	43	0.92
Other	21	0.80	9	1.37	3	0.46	5	0.76	4	0.61

Table A3.6 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-9										
Ala	437	0.89	121	0.99	118	0.96	114	0.93	84	0.68
Cys	74	1.03	28	1.56	11	0.61	18	1.01	17	0.95
Asp	307	0.90	45	0.53	63	0.74	139	1.64	60	0.71
Glu	309	0.81	61	0.64	83	0.87	116	1.22	49	0.52
Phe	195	0.84	80	1.38	41	0.71	36	0.62	38	0.65
Gly	663	1.54	112	1.04	211	1.96	31	0.29	309	2.88
His	132	0.97	44	1.30	18	0.53	33	0.97	37	1.09
Ile	201	0.61	65	0.79	29	0.35	31	0.38	76	0.93
Lys	237	0.75	38	0.48	88	1.12	71	0.90	40	0.51
Leu	429	0.81	157	1.19	101	0.76	62	0.47	109	0.82
Met	81	0.83	29	1.18	15	0.61	16	0.65	21	0.86
Asn	218	0.92	47	0.79	39	0.65	75	1.26	57	0.96
Pro	910	3.42	135	2.03	268	4.03	350	5.26	157	2.36
Gln	197	0.94	58	1.11	46	0.88	40	0.76	53	1.01
Arg	247	0.85	64	0.88	62	0.85	69	0.94	52	0.71
Ser	315	0.95	95	1.14	76	0.91	62	0.74	82	0.98
Thr	242	0.79	94	1.22	52	0.67	48	0.62	48	0.62
Val	255	0.62	78	0.76	47	0.46	48	0.47	82	0.80
Trp	88	1.09	32	1.58	17	0.84	18	0.89	21	1.04
Tyr	158	0.79	41	0.82	38	0.76	47	0.94	32	0.64
Other	21	0.74	5	0.71	6	0.85	5	0.71	5	0.71
type: o4-10										
Ala	229	0.58	77	0.78	9	0.09	67	0.68	76	0.77
Cys	64	1.12	19	1.32	1	0.07	18	1.25	26	1.81
Asp	219	0.80	45	0.66	19	0.28	119	1.75	36	0.53
Glu	198	0.65	60	0.79	9	0.12	90	1.18	39	0.51
Phe	122	0.66	51	1.10	2	0.04	35	0.75	34	0.73
Gly	1,249	3.63	117	1.36	1,036	12.02	14	0.16	82	0.95
His	95	0.87	31	1.14	3	0.11	36	1.32	25	0.92
Ile	198	0.75	50	0.76	1	0.02	67	1.02	80	1.22
Lys	178	0.70	68	1.08	8	0.13	62	0.98	40	0.63
Leu	326	0.77	116	1.09	5	0.05	84	0.79	121	1.14
Met	62	0.79	23	1.17	2	0.10	13	0.66	24	1.22
Asn	203	1.06	39	0.82	7	0.15	89	1.86	68	1.42
Pro	261	1.22	49	0.92	0	0.00	94	1.76	118	2.21
Gln	125	0.74	39	0.93	5	0.12	39	0.93	42	1.00
Arg	190	0.81	58	0.99	9	0.15	74	1.26	49	0.84
Ser	199	0.75	83	1.24	10	0.15	60	0.90	46	0.69
Thr	212	0.86	75	1.21	4	0.06	64	1.04	69	1.12
Val	243	0.74	59	0.72	4	0.05	57	0.70	123	1.50
Trp	78	1.20	35	2.16	3	0.18	26	1.60	14	0.86
Tyr	125	0.78	50	1.24	8	0.20	38	0.95	29	0.72
Other	8	0.35	2	0.35	1	0.18	0	0.00	5	0.88

Table A3.6 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-11										
Ala	244	0.68	74	0.82	63	0.70	6	0.07	101	1.12
Cys	44	0.84	20	1.52	9	0.68	0	0.00	15	1.14
Asp	116	0.46	36	0.58	47	0.75	4	0.06	29	0.46
Glu	155	0.56	61	0.87	69	0.99	3	0.04	22	0.32
Phe	186	1.09	53	1.24	61	1.43	7	0.16	65	1.52
Gly	207	0.66	114	1.44	41	0.52	2	0.03	50	0.63
His	75	0.75	21	0.84	24	0.96	2	0.08	28	1.12
Ile	155	0.64	48	0.80	47	0.78	1	0.02	59	0.98
Lys	135	0.58	37	0.64	61	1.05	2	0.03	35	0.60
Leu	283	0.73	72	0.74	94	0.97	3	0.03	114	1.17
Met	55	0.76	16	0.89	11	0.61	2	0.11	26	1.44
Asn	133	0.76	40	0.91	49	1.12	2	0.05	42	0.96
Pro	1,205	6.16	81	1.66	33	0.67	993	20.29	98	2.00
Gln	129	0.84	51	1.32	40	1.04	3	0.08	35	0.91
Arg	127	0.59	39	0.73	43	0.80	1	0.02	44	0.82
Ser	222	0.91	68	1.11	71	1.16	6	0.10	77	1.26
Thr	230	1.02	89	1.57	81	1.43	5	0.09	55	0.97
Val	240	0.80	79	1.05	73	0.97	1	0.01	87	1.16
Trp	61	1.03	12	0.81	35	2.35	0	0.00	14	0.94
Tyr	187	1.27	38	1.03	92	2.50	8	0.22	49	1.33
Other	15	0.72	2	0.38	7	1.35	0	0.00	6	1.15
type: o4-12										
Ala	170	0.67	61	0.96	12	0.19	46	0.72	51	0.80
Cys	43	1.16	13	1.40	3	0.32	14	1.51	13	1.40
Asp	140	0.79	40	0.91	16	0.36	36	0.82	48	1.09
Glu	142	0.72	34	0.69	16	0.32	52	1.05	40	0.81
Phe	128	1.06	51	1.69	14	0.46	38	1.26	25	0.83
Gly	655	2.94	42	0.75	521	9.34	44	0.79	48	0.86
His	60	0.85	21	1.19	9	0.51	19	1.08	11	0.62
Ile	110	0.65	30	0.70	0	0.00	32	0.75	48	1.13
Lys	159	0.97	35	0.86	20	0.49	73	1.78	31	0.76
Leu	159	0.58	66	0.96	11	0.16	21	0.31	61	0.89
Met	25	0.49	10	0.78	2	0.16	9	0.71	4	0.31
Asn	164	1.33	42	1.36	50	1.62	37	1.20	35	1.13
Pro	122	0.88	42	1.22	0	0.00	0	0.00	80	2.32
Gln	88	0.81	13	0.48	18	0.66	42	1.54	15	0.55
Arg	152	1.00	26	0.69	22	0.58	68	1.79	36	0.95
Ser	218	1.26	68	1.57	23	0.53	73	1.69	54	1.25
Thr	174	1.09	76	1.90	1	0.02	50	1.25	47	1.17
Val	128	0.60	23	0.43	0	0.00	49	0.92	56	1.06
Trp	31	0.74	9	0.86	2	0.19	5	0.48	15	1.43
Tyr	94	0.90	36	1.38	2	0.08	34	1.31	22	0.85
Other	6	0.41	4	1.09	0	0.00	0	0.00	2	0.55

Table A3.6 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-13										
Ala	98	0.85	23	0.79	18	0.62	2	0.07	55	1.90
Cys	20	1.18	5	1.18	3	0.71	0	0.00	12	2.84
Asp	57	0.71	8	0.40	23	1.15	12	0.60	14	0.70
Glu	42	0.47	12	0.54	15	0.67	2	0.09	13	0.58
Phe	45	0.82	16	1.17	16	1.17	2	0.15	11	0.80
Gly	358	3.53	22	0.87	5	0.20	291	11.48	40	1.58
His	20	0.62	8	1.00	3	0.37	1	0.12	8	1.00
Ile	107	1.38	41	2.12	57	2.95	0	0.00	9	0.47
Lys	27	0.36	4	0.22	9	0.48	1	0.05	13	0.70
Leu	100	0.80	38	1.22	37	1.19	1	0.03	24	0.77
Met	18	0.78	9	1.55	4	0.69	1	0.17	4	0.69
Asn	42	0.75	5	0.36	20	1.42	2	0.14	15	1.07
Pro	39	0.62	16	1.02	3	0.19	0	0.00	20	1.27
Gln	25	0.51	11	0.89	11	0.89	1	0.08	2	0.16
Arg	42	0.61	11	0.64	17	0.99	7	0.41	7	0.41
Ser	69	0.88	15	0.76	16	0.81	7	0.36	31	1.58
Thr	52	0.72	19	1.05	7	0.39	2	0.11	24	1.32
Val	139	1.44	62	2.57	59	2.45	1	0.04	17	0.71
Trp	14	0.73	2	0.42	6	1.26	0	0.00	6	1.26
Tyr	31	0.66	7	0.59	8	0.68	4	0.34	12	1.02
Other	3	0.45	3	1.80	0	0.00	0	0.00	0	0.00
type: o4-14										
Ala	113	0.97	26	0.89	32	1.10	37	1.27	18	0.62
Cys	20	1.18	4	0.94	4	0.94	10	2.36	2	0.47
Asp	99	1.23	8	0.40	51	2.53	12	0.60	28	1.39
Glu	70	0.78	14	0.62	16	0.71	13	0.58	27	1.20
Phe	74	1.34	15	1.09	15	1.09	30	2.18	14	1.02
Gly	93	0.91	33	1.29	10	0.39	28	1.10	22	0.86
His	47	1.46	13	1.61	10	1.24	16	1.99	8	0.99
Ile	50	0.64	22	1.13	1	0.05	12	0.62	15	0.77
Lys	48	0.64	14	0.75	10	0.53	11	0.59	13	0.70
Leu	89	0.71	46	1.47	15	0.48	10	0.32	18	0.57
Met	25	1.07	9	1.54	5	0.86	7	1.20	4	0.69
Asn	101	1.79	16	1.13	47	3.33	14	0.99	24	1.70
Pro	121	1.92	29	1.84	53	3.36	0	0.00	39	2.47
Gln	35	0.70	5	0.40	10	0.80	9	0.72	11	0.88
Arg	54	0.78	14	0.81	10	0.58	19	1.10	11	0.63
Ser	106	1.34	17	0.86	23	1.16	38	1.92	28	1.42
Thr	68	0.93	11	0.60	8	0.44	21	1.15	28	1.53
Val	57	0.59	23	0.95	3	0.12	12	0.50	19	0.78
Trp	26	1.35	8	1.67	4	0.83	11	2.29	3	0.63
Tyr	57	1.20	12	1.01	11	0.93	27	2.27	7	0.59
Other	3	0.45	0	0.00	1	0.60	2	1.19	0	0.00

Table A3.6 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-15										
Ala	58	0.60	23	0.94	7	0.29	8	0.33	20	0.82
Cys	12	0.85	4	1.12	2	0.56	4	1.12	2	0.56
Asp	66	0.98	35	2.07	10	0.59	11	0.65	10	0.59
Glu	42	0.56	11	0.58	8	0.42	11	0.58	12	0.64
Phe	45	0.98	11	0.95	9	0.78	15	1.30	10	0.87
Gly	357	4.19	18	0.84	191	8.94	130	6.09	18	0.84
His	29	1.08	11	1.63	1	0.15	7	1.04	10	1.48
Ile	16	0.25	6	0.37	0	0.00	0	0.00	10	0.61
Lys	42	0.67	7	0.45	4	0.26	12	0.77	19	1.21
Leu	48	0.46	26	0.99	3	0.11	7	0.27	12	0.46
Met	11	0.56	4	0.82	1	0.20	0	0.00	6	1.23
Asn	69	1.46	12	1.01	20	1.69	27	2.28	10	0.85
Pro	43	0.81	15	1.13	0	0.00	0	0.00	28	2.12
Gln	29	0.70	10	0.96	3	0.29	6	0.58	10	0.96
Arg	49	0.85	12	0.83	2	0.14	4	0.28	31	2.14
Ser	74	1.12	24	1.45	9	0.54	15	0.91	26	1.57
Thr	42	0.69	28	1.83	3	0.20	1	0.07	10	0.65
Val	38	0.47	11	0.54	1	0.05	1	0.05	25	1.23
Trp	20	1.25	6	1.49	1	0.25	6	1.49	7	1.74
Tyr	41	1.03	10	1.00	8	0.80	17	1.71	6	0.60
Other	5	0.89	0	0.00	1	0.71	2	1.42	2	1.42
type: o4-16										
Ala	46	0.65	15	0.85	5	0.28	12	0.68	14	0.79
Cys	12	1.17	4	1.55	1	0.39	2	0.78	5	1.94
Asp	36	0.74	9	0.73	8	0.65	10	0.82	9	0.73
Glu	34	0.62	5	0.37	9	0.66	18	1.32	2	0.15
Phe	35	1.05	15	1.79	2	0.24	11	1.32	7	0.84
Gly	152	2.46	22	1.42	94	6.07	12	0.77	24	1.55
His	16	0.82	6	1.23	5	1.02	3	0.61	2	0.41
Ile	36	0.76	10	0.85	5	0.42	4	0.34	17	1.44
Lys	40	0.88	10	0.88	15	1.32	9	0.79	6	0.53
Leu	46	0.61	15	0.79	6	0.31	7	0.37	18	0.94
Met	12	0.85	5	1.41	2	0.56	4	1.13	1	0.28
Asn	29	0.85	6	0.70	6	0.70	11	1.28	6	0.70
Pro	38	0.99	13	1.36	0	0.00	0	0.00	25	2.61
Gln	31	1.03	6	0.79	7	0.93	12	1.59	6	0.79
Arg	41	0.98	4	0.38	13	1.23	15	1.42	9	0.85
Ser	60	1.25	4	0.33	6	0.50	33	2.75	17	1.42
Thr	41	0.93	14	1.26	5	0.45	15	1.35	7	0.63
Val	59	1.01	20	1.36	9	0.61	13	0.88	17	1.15
Trp	12	1.03	5	1.72	1	0.34	4	1.37	2	0.69
Tyr	44	1.53	17	2.35	7	0.97	11	1.52	9	1.25
Other	4	0.99	1	0.98	0	0.00	0	0.00	3	2.95

Table A3.6 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-17										
Ala	41	0.57	7	0.39	1	0.06	20	1.12	13	0.73
Cys	6	0.58	1	0.38	0	0.00	3	1.15	2	0.77
Asp	26	0.53	9	0.73	0	0.00	4	0.32	13	1.05
Glu	19	0.34	13	0.94	0	0.00	1	0.07	5	0.36
Phe	26	0.77	12	1.42	0	0.00	7	0.83	7	0.83
Gly	51	0.82	29	1.85	0	0.00	7	0.45	15	0.96
His	9	0.46	1	0.20	0	0.00	2	0.40	6	1.21
Ile	31	0.65	8	0.67	1	0.08	12	1.01	10	0.84
Lys	30	0.65	10	0.87	0	0.00	10	0.87	10	0.87
Leu	59	0.77	18	0.94	0	0.00	23	1.20	18	0.94
Met	12	0.84	1	0.28	0	0.00	7	1.96	4	1.12
Asn	34	0.98	12	1.38	0	0.00	11	1.27	11	1.27
Pro	270	6.98	2	0.21	201	20.75	39	4.03	28	2.89
Gln	20	0.66	13	1.70	0	0.00	5	0.65	2	0.26
Arg	29	0.68	7	0.66	0	0.00	7	0.66	15	1.41
Ser	52	1.07	22	1.82	1	0.08	15	1.24	14	1.16
Thr	34	0.76	16	1.43	0	0.00	5	0.45	13	1.16
Val	48	0.81	11	0.74	4	0.27	18	1.21	15	1.01
Trp	7	0.60	2	0.68	0	0.00	4	1.36	1	0.34
Tyr	24	0.82	12	1.65	0	0.00	7	0.96	5	0.69
Other	4	0.97	2	1.95	0	0.00	1	0.97	1	0.97
type: o4-18										
Ala	45	0.69	17	1.05	0	0.00	20	1.23	8	0.49
Cys	8	0.85	1	0.42	0	0.00	2	0.85	5	2.11
Asp	29	0.65	6	0.53	1	0.09	15	1.34	7	0.62
Glu	31	0.62	10	0.80	1	0.08	14	1.11	6	0.48
Phe	35	1.14	8	1.04	0	0.00	17	2.22	10	1.30
Gly	47	0.83	29	2.04	0	0.00	9	0.63	9	0.63
His	16	0.89	3	0.67	1	0.22	6	1.34	6	1.34
Ile	17	0.39	1	0.09	0	0.00	7	0.65	9	0.83
Lys	45	1.08	28	2.69	0	0.00	9	0.86	8	0.77
Leu	33	0.47	10	0.57	0	0.00	13	0.74	10	0.57
Met	9	0.69	3	0.92	0	0.00	2	0.62	4	1.23
Asn	17	0.54	3	0.38	0	0.00	11	1.40	3	0.38
Pro	231	6.56	21	2.39	180	20.45	0	0.00	30	3.41
Gln	13	0.47	2	0.29	0	0.00	5	0.72	6	0.86
Arg	26	0.67	4	0.41	0	0.00	13	1.35	9	0.93
Ser	39	0.89	11	1.00	3	0.27	14	1.27	11	1.00
Thr	27	0.66	6	0.59	2	0.20	2	0.20	17	1.67
Val	35	0.65	6	0.44	0	0.00	12	0.89	17	1.26
Trp	18	1.68	8	2.99	0	0.00	5	1.87	5	1.87
Tyr	33	1.25	11	1.66	1	0.15	12	1.81	9	1.36
Other	2	0.54	1	1.07	0	0.00	1	1.07	0	0.00

Table A3.6 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 4 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o4-19										
Ala	33	0.65	11	0.86	12	0.94	0	0.00	10	0.78
Cys	6	0.81	3	1.61	2	1.07	1	0.54	0	0.00
Asp	40	1.13	14	1.58	16	1.81	2	0.23	8	0.90
Glu	33	0.83	10	1.01	10	1.01	2	0.20	11	1.11
Phe	11	0.46	2	0.33	5	0.83	0	0.00	4	0.66
Gly	168	3.75	15	1.34	4	0.36	138	12.32	11	0.98
His	14	0.99	6	1.70	5	1.41	0	0.00	3	0.85
Ile	16	0.47	3	0.35	5	0.58	0	0.00	8	0.94
Lys	30	0.91	9	1.10	10	1.22	0	0.00	11	1.34
Leu	27	0.49	5	0.36	13	0.94	1	0.07	8	0.58
Met	7	0.68	1	0.39	4	1.56	0	0.00	2	0.78
Asn	30	1.21	6	0.97	13	2.09	0	0.00	11	1.77
Pro	23	0.83	9	1.30	2	0.29	0	0.00	12	1.73
Gln	20	0.92	9	1.65	6	1.10	0	0.00	5	0.91
Arg	23	0.76	4	0.53	11	1.44	3	0.39	5	0.66
Ser	37	1.07	20	2.30	6	0.69	1	0.12	10	1.15
Thr	29	0.90	11	1.37	10	1.24	1	0.12	7	0.87
Val	23	0.54	3	0.28	8	0.75	0	0.00	12	1.13
Trp	8	0.95	4	1.90	1	0.47	0	0.00	3	1.42
Tyr	14	0.67	3	0.57	4	0.77	0	0.00	7	1.34
Other	4	1.36	1	1.36	2	2.72	0	0.00	1	1.36
type: o4-20										
Ala	21	0.37	9	0.63	0	0.00	3	0.21	9	0.63
Cys	6	0.72	1	0.48	0	0.00	3	1.44	2	0.96
Asp	32	0.81	7	0.71	3	0.30	15	1.52	7	0.71
Glu	36	0.82	17	1.54	1	0.09	7	0.63	11	1.00
Phe	27	1.00	4	0.59	0	0.00	8	1.19	15	2.23
Gly	158	3.16	6	0.48	142	11.37	1	0.08	9	0.72
His	13	0.82	2	0.51	1	0.25	6	1.52	4	1.01
Ile	35	0.92	4	0.42	0	0.00	22	2.31	9	0.95
Lys	25	0.68	8	0.87	2	0.22	9	0.98	6	0.66
Leu	37	0.60	5	0.33	0	0.00	16	1.04	16	1.04
Met	9	0.79	2	0.70	0	0.00	6	2.10	1	0.35
Asn	44	1.59	10	1.45	5	0.72	18	2.60	11	1.59
Pro	53	1.71	34	4.40	0	0.00	0	0.00	19	2.46
Gln	17	0.70	7	1.15	0	0.00	9	1.48	1	0.16
Arg	26	0.77	11	1.30	2	0.24	11	1.30	2	0.24
Ser	35	0.90	15	1.55	4	0.41	10	1.03	6	0.62
Thr	34	0.95	17	1.90	2	0.22	2	0.22	13	1.45
Val	29	0.61	5	0.42	1	0.08	7	0.59	16	1.35
Trp	12	1.28	1	0.43	1	0.43	6	2.55	4	1.70
Tyr	15	0.64	1	0.17	2	0.34	7	1.20	5	0.86
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

OPEN 5 TURNS

Table A3.7. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 5 turn structures. Turn types o5-23 and o5-48 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		ψ_{i+1}		φ_{i+1}		ω_{i+1}		ψ_{i+2}		φ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o5-23	310	177.4	5.8	-89.1	26.0	150.3	15.6	176.6	6.3	-108.8	26.3	132.6	16.0
o5-23	*213*	*177.3*	*5.8*	*-87.4*	*25.9*	*149.5*	*16.2*	*176.7*	*6.1*	*-110.6*	*27.6*	*131.6*	*16.3*
o5-24	287	180.0	4.8	-143.4	13.7	166.2	12.6	175.3	5.2	64.9	13.9	-144.1	14.7
o5-24	*200*	*-179.8*	*5.1*	*-143.0*	*13.9*	*166.3*	*12.4*	*175.3*	*5.4*	*65.1*	*13.6*	*-144.8*	*14.0*
o5-25	194	179.7	5.2	-91.4	23.0	153.1	17.6	174.6	5.7	-61.4	14.6	144.4	7.1
o5-25	*136*	*179.8*	*5.7*	*-91.8*	*23.3*	*152.8*	*18.4*	*175.0*	*6.4*	*-61.9*	*16.3*	*144.1*	*7.7*
o5-26	148	-178.9	4.2	89.9	14.0	-14.5	14.6	179.1	5.6	-112.8	17.9	-39.0	15.7
o5-26	*116*	*-178.5*	*4.4*	*89.0*	*14.2*	*-13.6*	*13.4*	*179.5*	*4.0*	*-115.3*	*17.7*	*-38.9*	*16.6*
o5-27	158	-179.7	5.4	-82.1	16.4	-27.0	15.3	179.1	5.3	-143.1	15.6	151.3	17.6
o5-27	*122*	*-179.3*	*5.8*	*-81.5*	*15.8*	*-26.5*	*15.1*	*179.0*	*5.1*	*-141.9*	*17.7*	*153.3*	*15.9*
o5-28	139	-179.1	5.6	75.9	16.7	9.6	23.0	178.8	5.5	-157.2	14.5	166.9	11.2
o5-28	*100*	*-178.3*	*5.8*	*75.9*	*16.1*	*9.4*	*24.4*	*177.9*	*6.0*	*-159.5*	*16.8*	*167.5*	*12.1*
o5-29	144	177.3	6.3	-132.9	12.2	120.0	11.1	-2.1	4.6	-69.1	7.9	161.1	10.8
o5-29	*112*	*177.3*	*5.8*	*-133.0*	*11.1*	*119.0*	*11.6*	*-1.8*	*4.3*	*-69.0*	*8.4*	*160.9*	*11.0*
o5-30	128	178.7	5.6	96.8	18.4	175.9	21.9	-178.8	4.6	-67.9	18.8	137.7	12.2
o5-30	*101*	*178.2*	*5.5*	*98.8*	*17.5*	*176.0*	*20.6*	*-178.8*	*5.6*	*-65.3*	*16.8*	*137.3*	*11.9*
o5-31	123	-179.8	5.1	-87.9	9.2	59.5	18.5	-179.1	6.8	-149.5	16.5	173.5	13.1
o5-31	*84*	*179.0*	*5.3*	*-87.3*	*8.0*	*61.7*	*18.2*	*-178.7*	*6.2*	*-149.5*	*16.0*	*175.2*	*14.0*
o5-32	99	179.5	5.6	-114.5	16.0	-79.6	22.4	178.5	8.4	-91.6	19.0	-18.5	14.9
o5-32	*59*	*-179.4*	*5.6*	*-116.7*	*16.4*	*-79.1*	*21.4*	*179.0*	*7.5*	*-90.9*	*16.4*	*-23.6*	*13.5*
o5-33	109	178.3	5.0	75.9	13.3	15.2	14.2	-178.6	4.0	-111.5	16.0	5.4	11.2
o5-33	*64*	*178.9*	*4.7*	*73.6*	*13.9*	*18.6*	*14.3*	*-179.3*	*4.6*	*-114.0*	*15.8*	*6.0*	*11.6*
o5-34	100	-178.6	6.2	-106.9	14.0	5.8	10.2	-178.2	3.8	93.5	15.5	-14.5	16.1
o5-34	*78*	*-178.0*	*5.4*	*-108.0*	*15.5*	*6.1*	*10.9*	*-178.3*	*4.1*	*93.4*	*15.1*	*-14.4*	*15.3*
o5-35	70	179.2	5.7	90.9	13.0	-173.4	15.8	-174.9	4.9	-79.4	12.9	-20.6	14.2
o5-35	*49*	*178.9*	*6.3*	*90.5*	*12.2*	*-172.2*	*14.6*	*-174.9*	*4.5*	*-76.2*	*9.5*	*-23.6*	*13.1*
o5-36	105	2.7	4.5	-90.6	8.5	1.3	13.2	-175.6	5.5	-58.3	6.9	-36.6	11.3
o5-36	*72*	*2.7*	*4.4*	*-91.6*	*7.6*	*0.0*	*14.6*	*-175.3*	*5.4*	*-58.9*	*6.5*	*-36.9*	*12.0*
o5-37	72	-178.4	5.9	-90.1	13.0	3.4	21.6	179.4	4.4	-108.6	13.1	-131.1	15.3
o5-37	*56*	*-179.2*	*5.0*	*-93.3*	*13.2*	*3.3*	*18.7*	*-179.8*	*3.9*	*-109.1*	*12.6*	*-132.5*	*14.5*
o5-38	81	179.6	3.4	77.3	12.9	10.5	18.0	178.2	4.4	-127.1	22.0	172.7	10.9
o5-38	*68*	*179.6*	*3.1*	*77.7*	*12.7*	*9.8*	*19.7*	*177.5*	*4.3*	*-127.5*	*23.1*	*173.6*	*12.1*
o5-39	63	177.1	4.9	-143.9	19.3	-126.0	13.6	174.9	5.6	-63.0	9.3	-34.5	8.4
o5-39	*42*	*178.1*	*5.0*	*-142.7*	*21.0*	*-123.3*	*14.9*	*175.8*	*6.5*	*-62.9*	*7.0*	*-36.5*	*9.0*
o5-40	65	-179.8	3.6	87.5	11.4	-7.6	12.5	-179.7	3.9	-113.9	12.7	-60.7	13.2
o5-40	*50*	*-179.0*	*3.9*	*85.0*	*10.8*	*-7.1*	*12.7*	*179.1*	*5.0*	*-112.6*	*13.2*	*-58.4*	*10.4*
o5-41	64	179.3	6.1	-51.8	7.0	139.5	5.9	6.9	5.0	-88.3	6.7	10.5	7.1
o5-41	*42*	*179.4*	*5.2*	*-52.4*	*7.2*	*139.8*	*6.1*	*6.2*	*5.1*	*-89.8*	*6.7*	*12.1*	*7.4*
o5-42	49	1.8	3.7	-89.7	5.7	0.9	11.3	-177.6	5.0	-66.6	10.3	-36.4	9.6
o5-42	*30*	*1.4*	*3.0*	*-88.7*	*6.9*	*-0.4*	*12.3*	*-177.6*	*6.0*	*-68.1*	*10.9*	*-33.7*	*9.5*
o5-43	57	-1.9	3.9	-73.9	8.0	156.9	16.2	-176.9	3.4	-58.7	7.4	-32.2	9.4
o5-43	*47*	*-1.7*	*4.4*	*-75.4*	*8.1*	*155.3*	*18.1*	*-177.0*	*4.1*	*-58.0*	*7.1*	*-32.9*	*9.1*
o5-44	48	176.0	4.2	-55.9	7.8	142.1	6.8	3.5	4.1	-93.5	5.3	13.6	12.8
o5-44	*30*	*175.6*	*5.1*	*-56.7*	*7.4*	*141.0*	*5.1*	*3.2*	*4.1*	*-94.3*	*5.7*	*14.3*	*12.5*
o5-45	42	-179.3	4.6	-124.2	17.2	138.5	11.5	2.4	3.9	-81.7	7.7	-10.6	10.4
o5-45	*30*	*179.4*	*4.4*	*-123.6*	*19.2*	*140.0*	*8.0*	*1.4*	*3.8*	*-79.7*	*8.6*	*-12.0*	*12.4*
o5-46	56	178.1	5.9	53.1	5.1	37.9	7.8	179.6	3.1	57.3	5.8	27.5	9.1
o5-46	*43*	*178.8*	*6.6*	*53.0*	*5.5*	*38.6*	*7.7*	*179.2*	*4.6*	*57.5*	*4.8*	*28.5*	*9.3*
o5-47	33	179.1	5.0	75.2	11.1	-167.1	14.3	-180.0	3.3	57.2	6.5	-137.5	15.8
o5-47	*22*	*179.9*	*4.9*	*75.5*	*9.8*	*-162.3*	*13.3*	*179.5*	*7.2*	*55.8*	*5.5*	*-138.9*	*15.9*
o5-48	41	178.2	4.4	-60.7	5.6	128.5	5.9	-178.5	2.7	84.4	7.9	1.0	8.4
o5-48	*31*	*178.6*	*4.6*	*-62.3*	*6.4*	*128.7*	*8.4*	*-178.8*	*3.1*	*85.8*	*10.4*	*1.6*	*8.5*

Table A3.7 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and Cα Cα distances of the classified turn types for the training and test set of open 5 turn structures. Turn types o5-23 and o5-48 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+2}		ψ_{i+3}		φ_{i+3}		ω_{i+3}		Cα-Cα [Å]	
		∅	±	∅	±	∅	∅	∅	±	∅	±
o5-23	310	-1.5	4.6	-73.9	8.9	151.7	13.6	-179.3	7.4	6.9	0.8
o5-23	*213*	*-1.8*	*5.0*	*-73.0*	*8.2*	*152.4*	*12.8*	*-179.7*	*5.0*	*6.9*	*0.8*
o5-24	287	-176.5	4.0	-61.1	7.0	-33.1	9.5	179.7	3.4	6.4	0.8
o5-24	*200*	*-176.4*	*4.3*	*-61.6*	*6.7*	*-32.8*	*8.8*	*179.9*	*3.7*	*6.4*	*0.8*
o5-25	194	4.9	4.8	-88.4	7.2	4.9	14.3	179.9	7.5	6.4	0.7
o5-25	*136*	*4.5*	*5.1*	*-87.4*	*8.2*	*4.0*	*15.0*	*-179.5*	*6.2*	*6.4*	*0.7*
o5-26	148	-179.9	4.8	-134.1	18.8	145.1	20.9	177.5	6.2	7.1	0.6
o5-26	*116*	*179.3*	*4.6*	*-134.2*	*20.8*	*148.2*	*19.4*	*176.3*	*6.4*	*7.1*	*0.6*
o5-27	158	177.0	6.0	74.2	15.4	-155.9	17.8	-177.3	4.3	7.2	0.6
o5-27	*122*	*176.7*	*5.7*	*75.3*	*17.2*	*-155.8*	*18.4*	*-177.5*	*5.1*	*7.3*	*0.6*
o5-28	139	177.0	5.9	-69.0	12.0	126.7	20.3	-179.6	6.1	7.5	0.4
o5-28	*100*	*177.1*	*6.3*	*-69.1*	*11.9*	*127.1*	*20.3*	*-179.2*	*6.0*	*7.5*	*0.4*
o5-29	144	179.8	4.5	-64.2	7.7	144.7	11.5	177.9	5.4	7.0	0.5
o5-29	*112*	*179.4*	*4.7*	*-63.0*	*7.8*	*143.8*	*11.7*	*178.5*	*5.6*	*6.9*	*0.5*
o5-30	128	178.4	4.6	69.0	14.2	18.6	20.9	-179.9	4.1	5.9	0.8
o5-30	*101*	*178.7*	*4.4*	*69.2*	*14.1*	*17.6*	*20.4*	*179.6*	*4.2*	*5.9*	*0.8*
o5-31	123	-179.2	5.6	-65.1	12.1	-25.9	18.3	179.6	3.7	7.5	0.5
o5-31	*84*	*-179.2*	*4.9*	*-66.1*	*13.2*	*-25.7*	*19.2*	*179.5*	*4.0*	*7.4*	*0.6*
o5-32	99	-179.9	6.2	-109.6	22.7	124.2	18.8	178.9	7.2	6.1	1.1
o5-32	*59*	*-178.9*	*6.0*	*-109.0*	*24.6*	*124.7*	*19.9*	*178.8*	*6.5*	*6.1*	*1.1*
o5-33	109	179.4	4.3	82.0	15.6	15.9	15.5	-179.4	3.8	7.4	0.4
o5-33	*64*	*-179.9*	*4.5*	*80.0*	*16.1*	*15.4*	*15.2*	*-179.2*	*3.7*	*7.3*	*0.5*
o5-34	100	-179.5	3.4	-120.4	12.8	-37.5	17.3	-178.6	4.3	6.3	0.9
o5-34	*78*	*-180.0*	*4.6*	*-120.0*	*14.7*	*-37.8*	*18.1*	*-178.9*	*4.6*	*6.3*	*0.9*
o5-35	70	-179.2	5.0	-122.7	15.6	102.1	30.3	-178.5	6.6	6.5	0.9
o5-35	*49*	*-177.9*	*5.5*	*-123.1*	*15.2*	*108.5*	*30.8*	*-178.9*	*6.9*	*6.5*	*0.8*
o5-36	105	-179.6	3.9	-68.1	13.7	-25.2	17.8	-179.3	4.5	6.6	0.6
o5-36	*72*	*-179.5*	*3.5*	*-68.6*	*14.2*	*-24.4*	*17.5*	*-179.7*	*4.3*	*6.6*	*0.6*
o5-37	72	-176.4	5.6	-65.6	13.9	-34.3	10.7	-178.6	5.7	5.9	0.9
o5-37	*56*	*-177.0*	*4.4*	*-64.1*	*13.2*	*-38.0*	*11.2*	*-179.0*	*4.5*	*5.9*	*0.9*
o5-38	81	180.0	4.4	-64.3	11.2	-21.7	13.8	179.7	3.9	7.5	0.3
o5-38	*68*	*-179.9*	*4.8*	*-64.9*	*12.4*	*-21.2*	*14.2*	*178.9*	*3.6*	*7.5*	*0.4*
o5-39	63	174.3	5.7	-123.5	13.5	168.5	11.5	176.8	5.9	4.9	0.8
o5-39	*42*	*175.6*	*6.8*	*-122.6*	*16.8*	*170.4*	*13.4*	*177.2*	*5.3*	*4.9*	*0.9*
o5-40	65	179.9	3.9	-88.9	18.4	-21.2	14.5	178.9	5.4	5.9	0.7
o5-40	*50*	*-179.5*	*4.2*	*-91.6*	*17.8*	*-18.4*	*14.4*	*179.7*	*5.1*	*5.9*	*0.7*
o5-41	64	177.6	5.4	-75.5	12.2	148.0	13.6	178.1	6.5	7.2	0.6
o5-41	*42*	*179.3*	*5.5*	*-76.2*	*12.4*	*149.3*	*14.9*	*178.5*	*5.0*	*7.2*	*0.5*
o5-42	49	177.8	6.4	-122.7	19.8	134.1	21.4	179.2	5.3	6.1	0.8
o5-42	*30*	*178.0*	*6.6*	*-120.4*	*19.1*	*133.3*	*18.0*	*-179.3*	*4.0*	*6.2*	*0.6*
o5-43	57	-179.0	3.6	-66.9	11.0	-22.7	14.2	179.1	3.8	6.9	0.6
o5-43	*47*	*-177.8*	*7.7*	*-66.1*	*10.4*	*-24.1*	*13.2*	*178.3*	*3.9*	*7.0*	*0.7*
o5-44	48	-177.2	4.3	-73.0	17.9	-19.2	22.2	178.2	4.9	6.0	0.8
o5-44	*30*	*-178.1*	*5.9*	*-75.7*	*17.0*	*-19.4*	*21.3*	*177.8*	*5.2*	*5.9*	*0.7*
o5-45	42	178.3	5.2	-136.2	16.3	146.4	22.1	178.6	6.4	4.7	0.6
o5-45	*30*	*178.1*	*5.0*	*-139.3*	*16.0*	*149.1*	*21.4*	*176.9*	*6.1*	*4.7*	*0.7*
o5-46	56	179.1	3.7	78.6	11.4	10.5	10.9	-179.2	4.7	7.5	0.4
o5-46	*43*	*179.8*	*3.6*	*77.2*	*12.3*	*11.7*	*11.3*	*179.5*	*4.4*	*7.4*	*0.3*
o5-47	33	-178.3	3.1	-74.3	14.0	-16.5	18.9	-179.1	4.6	6.4	0.7
o5-47	*22*	*-177.5*	*4.9*	*-72.7*	*12.9*	*-16.8*	*16.0*	*-178.9*	*4.6*	*6.4*	*0.7*
o5-48	41	177.9	5.9	81.8	16.9	179.2	28.5	-178.6	4.7	6.5	1.0
o5-48	*31*	*178.4*	*6.5*	*77.3*	*19.1*	*-173.2*	*29.0*	*-178.1*	*3.2*	*6.8*	*0.9*

Table A3.8. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 5 turn structures. Turn types o5-49 and o5-56 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		ψ_{i+1}		φ_{i+1}		ω_{i+1}		ψ_{i+2}		φ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o5-49	32	179.0	3.7	-64.4	6.2	-23.6	6.2	-178.9	3.4	-104.5	10.2	5.7	13.6
o5-49	*26*	*178.7*	*3.7*	*-64.6*	*6.2*	*-23.5*	*9.2*	*-178.4*	*3.7*	*-106.2*	*9.4*	*8.8*	*11.2*
o5-50	24	-179.9	5.6	-107.7	21.0	13.8	13.3	179.9	8.8	103.5	13.1	-4.5	11.5
o5-50	*11*	*179.5*	*5.2*	*-97.5*	*22.3*	*7.0*	*14.8*	*-177.9*	*6.0*	*104.6*	*13.4*	*-5.7*	*7.2*
o5-51	31	180.0	4.6	57.3	4.1	-126.4	9.1	-180.0	1.8	-80.6	11.4	-2.7	10.1
o5-51	*25*	*179.8*	*5.6*	*57.1*	*6.2*	*-123.8*	*8.4*	*179.9*	*2.6*	*-79.2*	*11.1*	*-3.4*	*12.3*
o5-52	25	179.6	2.5	-103.2	10.1	12.3	9.9	179.9	4.4	96.7	17.8	-79.7	17.2
o5-52	*14*	*179.1*	*2.1*	*-103.8*	*11.1*	*12.5*	*8.0*	*179.7*	*5.2*	*99.2*	*22.1*	*-80.4*	*20.0*
o5-53	22	-175.5	6.3	84.0	11.7	-59.7	7.0	178.0	5.1	-62.1	6.8	-41.3	8.7
o5-53	*15*	*-178.7*	*4.6*	*76.3*	*11.3*	*-57.6*	*6.5*	*176.1*	*5.9*	*-56.3*	*14.1*	*-45.7*	*16.0*
o5-54	17	175.7	7.5	-128.7	8.5	99.5	16.5	-176.3	5.5	-52.8	9.1	138.8	3.9
o5-54	*11*	*178.1*	*6.8*	*-132.6*	*9.4*	*101.4*	*20.6*	*-177.1*	*5.0*	*-52.8*	*8.0*	*139.3*	*4.2*
o5-55	16	-180.0	5.3	-97.2	11.1	-28.5	18.3	178.2	4.8	-141.9	9.3	125.8	10.0
o5-55	*12*	*-179.3*	*5.7*	*-101.6*	*17.0*	*-28.7*	*17.3*	*176.0*	*4.9*	*-143.2*	*17.9*	*128.9*	*12.5*
o5-56	13	177.9	7.0	-134.4	5.2	105.2	19.4	0.6	4.4	-74.0	11.7	166.0	5.4
o5-56	*6*	*-179.4*	*8.7*	*-137.3*	*5.9*	*105.1*	*19.6*	*0.1*	*4.2*	*-75.2*	*9.7*	*163.5*	*7.4*

type	|turns|	ω_{i+2}		ψ_{i+3}		φ_{i+3}		ω_{i+3}		Cα-Cα[Å]	
		∅	±	∅	±	∅	∅	∅	±	∅	±
o5-49	*26*	*178.9*	*4.7*	*-76.0*	*21.3*	*146.6*	*7.3*	*1.8*	*8.0*	*7.7*	*0.2*
o5-49	*26*	*178.9*	*4.7*	*-76.0*	*21.3*	*146.6*	*7.3*	*1.8*	*8.0*	*7.7*	*0.2*
o5-50	24	174.3	5.7	-139.7	12.9	164.5	11.2	179.4	5.1	7.3	1.1
o5-50	*11*	*171.6*	*5.2*	*-140.0*	*13.2*	*167.2*	*13.2*	*-179.9*	*5.5*	*7.1*	*1.2*
o5-51	31	176.1	5.5	91.5	9.2	8.6	9.2	-178.6	4.5	7.5	0.3
o5-51	*25*	*175.9*	*4.8*	*90.9*	*10.4*	*8.5*	*7.9*	*-178.7*	*4.7*	*7.4*	*0.4*
o5-52	25	178.8	3.3	-81.0	8.9	-12.4	17.1	-179.9	4.5	6.4	0.5
o5-52	*14*	*177.4*	*3.3*	*-77.5*	*10.0*	*-14.4*	*12.8*	*178.6*	*2.8*	*6.3*	*0.3*
o5-53	22	-179.7	3.5	-63.6	7.5	-35.3	15.0	178.9	3.5	7.0	0.7
o5-53	*15*	*179.8*	*3.1*	*-64.0*	*4.9*	*-38.5*	*17.3*	*178.6*	*3.0*	*7.0*	*0.7*
o5-54	17	4.3	4.5	-90.5	3.9	12.6	15.9	-177.3	6.8	6.6	0.8
o5-54	*11*	*4.7*	*5.8*	*-91.2*	*3.5*	*19.1*	*7.9*	*-179.1*	*6.2*	*6.5*	*0.8*
o5-55	16	-6.4	5.0	-66.5	7.5	153.0	13.5	-179.1	2.4	6.7	1.0
o5-55	*12*	*-3.6*	*5.6*	*-71.8*	*8.3*	*152.8*	*17.1*	*-178.4*	*2.5*	*6.9*	*1.0*
o5-56	13	-178.0	4.0	-57.3	7.4	-35.3	5.4	-178.5	3.1	7.2	0.5
o5-56	*6*	*-176.4*	*4.0*	*-60.5*	*5.5*	*-34.8*	*4.9*	*-179.5*	*3.4*	*7.4*	*0.4*

Table A3.9. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*		*position 4*		*position 5*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-1												
Ala	21,318	1.19	4,146	1.16	4,639	1.29	5,021	1.40	4,276	1.19	3,236	0.90
Cys	2,376	0.91	416	0.80	476	0.91	413	0.79	484	0.93	587	1.12
Asp	11,746	0.95	3,227	1.30	1,786	0.72	2,552	1.03	2,596	1.05	1,585	0.64
Glu	15,269	1.10	3,407	1.23	3,104	1.12	3,897	1.41	3,018	1.09	1,843	0.67
Phe	8,813	1.04	1,602	0.95	1,936	1.14	1,443	0.85	1,903	1.12	1,929	1.14
Gly	9,427	0.60	2,126	0.68	1,299	0.41	1,564	0.50	1,727	0.55	2,711	0.86
His	5,229	1.06	1,099	1.11	804	0.81	987	1.00	1,113	1.12	1,226	1.24
Ile	10,238	0.86	1,870	0.78	2,834	1.18	1,855	0.78	1,247	0.52	2,432	1.02
Lys	13,129	1.14	2,353	1.02	2,181	0.95	2,773	1.21	3,073	1.34	2,749	1.20
Leu	24,445	1.27	3,934	1.02	5,862	1.52	5,041	1.31	4,679	1.21	4,929	1.28
Met	4,120	1.15	725	1.01	906	1.26	847	1.18	809	1.13	833	1.16
Asn	8,593	0.99	1,917	1.10	1,064	0.61	1,581	0.91	2,291	1.32	1,740	1.00
Pro	7,544	0.78	2,078	1.07	2,247	1.16	828	0.43	891	0.46	1,500	0.77
Gln	9,112	1.19	1,617	1.06	1,447	0.94	1,862	1.22	2,219	1.45	1,967	1.28
Arg	11,744	1.10	2,251	1.06	2,090	0.98	2,463	1.16	2,444	1.15	2,496	1.17
Ser	13,454	1.11	2,645	1.09	1,991	0.82	2,922	1.20	3,243	1.33	2,653	1.09
Thr	9,202	0.82	1,920	0.85	1,524	0.68	1,549	0.69	1,972	0.88	2,237	0.99
Val	11,054	0.74	2,105	0.71	2,897	0.97	1,932	0.65	1,423	0.48	2,697	0.90
Trp	3,325	1.13	681	1.15	865	1.47	647	1.10	518	0.88	614	1.04
Tyr	7,281	1.00	1,383	0.95	1,509	1.03	1,289	0.88	1,563	1.07	1,537	1.05
Other	1,131	1.10	208	1.01	249	1.21	244	1.18	221	1.07	209	1.01
type: o5-2												
Ala	10,622	0.87	1,999	0.82	1,054	0.43	2,746	1.12	2,637	1.08	2,186	0.89
Cys	1,543	0.87	460	1.29	359	1.01	170	0.48	153	0.43	401	1.12
Asp	14,361	1.70	910	0.54	4,328	2.55	1,848	1.09	4,071	2.40	3,204	1.89
Glu	12,929	1.37	1,132	0.60	931	0.49	2,854	1.51	4,595	2.43	3,417	1.80
Phe	4,362	0.75	1,480	1.28	330	0.29	788	0.68	710	0.61	1,054	0.91
Gly	9,403	0.88	1,865	0.87	1,860	0.87	838	0.39	1,278	0.60	3,562	1.66
His	3,467	1.03	559	0.83	1,037	1.53	402	0.59	710	1.05	759	1.12
Ile	4,902	0.60	2,304	1.41	178	0.11	1,015	0.62	560	0.34	845	0.52
Lys	6,445	0.82	1,281	0.82	831	0.53	1,699	1.08	1,641	1.04	993	0.63
Leu	9,342	0.71	3,719	1.41	731	0.28	2,057	0.78	1,182	0.45	1,653	0.63
Met	1,946	0.80	777	1.59	146	0.30	341	0.70	259	0.53	423	0.86
Asn	6,829	1.15	823	0.69	2,359	1.99	691	0.58	1,775	1.49	1,181	0.99
Pro	10,226	1.54	1,512	1.14	2,806	2.11	5,464	4.12	390	0.29	54	0.04
Gln	5,072	0.97	842	0.80	480	0.46	856	0.82	1,280	1.22	1,614	1.54
Arg	5,596	0.77	1,218	0.84	856	0.59	1,335	0.92	1,307	0.90	880	0.60
Ser	12,532	1.51	1,276	0.77	5,284	3.18	1,827	1.10	2,457	1.48	1,688	1.02
Thr	9,863	1.28	1,717	1.12	4,073	2.65	983	0.64	1,488	0.97	1,602	1.04
Val	6,851	0.67	2,868	1.41	285	0.14	1,441	0.71	722	0.35	1,535	0.75
Trp	1,807	0.90	479	1.19	138	0.34	419	1.04	421	1.04	350	0.87
Tyr	3,792	0.76	1,044	1.05	389	0.39	603	0.60	791	0.79	965	0.97
Other	610	0.87	235	1.67	45	0.32	123	0.87	73	0.52	134	0.95

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-3												
Ala	2,210	0.87	519	1.02	622	1.22	393	0.77	385	0.75	291	0.57
Cys	373	1.00	76	1.02	37	0.50	69	0.93	136	1.83	55	0.74
Asp	2,534	1.43	492	1.39	468	1.32	613	1.73	412	1.17	549	1.55
Glu	1,912	0.97	259	0.66	650	1.65	455	1.15	222	0.56	326	0.83
Phe	1,113	0.92	212	0.88	132	0.55	249	1.03	377	1.56	143	0.59
Gly	1,344	0.60	384	0.86	181	0.40	188	0.42	321	0.72	270	0.60
His	959	1.36	176	1.25	119	0.84	252	1.78	295	2.09	117	0.83
Ile	1,050	0.62	225	0.66	167	0.49	130	0.38	286	0.84	242	0.71
Lys	1,875	1.14	287	0.88	570	1.74	417	1.27	232	0.71	369	1.13
Leu	2,177	0.79	618	1.12	405	0.74	415	0.75	438	0.80	301	0.55
Met	437	0.86	101	0.99	70	0.69	92	0.90	110	1.08	64	0.63
Asn	1,836	1.48	373	1.51	211	0.85	541	2.18	425	1.72	286	1.15
Pro	2,028	1.47	317	1.15	564	2.04	12	0.04	0	0.00	1,135	4.10
Gln	1,137	1.04	170	0.78	292	1.34	294	1.35	194	0.89	187	0.86
Arg	1,591	1.05	298	0.98	365	1.20	392	1.29	260	0.86	276	0.91
Ser	2,424	1.40	542	1.56	487	1.41	441	1.27	524	1.51	430	1.24
Thr	1,845	1.15	336	1.05	260	0.81	366	1.14	470	1.47	413	1.29
Val	1,311	0.62	247	0.58	160	0.38	191	0.45	419	0.99	294	0.69
Trp	381	0.91	83	0.99	64	0.76	92	1.09	87	1.03	55	0.65
Tyr	1,068	1.03	206	0.99	102	0.49	315	1.51	320	1.54	125	0.60
Other	120	0.82	24	0.82	19	0.65	28	0.95	32	1.09	17	0.58
type: o5-4												
Ala	1,675	0.84	310	0.77	370	0.92	470	1.17	83	0.21	442	1.10
Cys	254	0.87	97	1.66	35	0.60	17	0.29	17	0.29	88	1.51
Asp	1,241	0.90	140	0.51	123	0.44	278	1.00	212	0.77	488	1.76
Glu	1,419	0.92	122	0.39	338	1.09	489	1.58	88	0.28	382	1.23
Phe	688	0.73	248	1.31	99	0.52	91	0.48	78	0.41	172	0.91
Gly	4,432	2.53	345	0.98	338	0.96	94	0.27	3,181	9.08	474	1.35
His	475	0.86	116	1.05	85	0.77	71	0.64	89	0.80	114	1.03
Ile	931	0.70	469	1.75	206	0.77	119	0.45	4	0.01	133	0.50
Lys	1,268	0.99	157	0.61	460	1.79	381	1.48	77	0.30	193	0.75
Leu	1,364	0.63	682	1.58	305	0.71	175	0.41	44	0.10	158	0.37
Met	317	0.79	93	1.16	59	0.74	47	0.59	25	0.31	93	1.16
Asn	922	0.95	119	0.61	129	0.66	112	0.58	426	2.19	136	0.70
Pro	2,223	2.05	277	1.28	665	3.06	1,262	5.82	0	0.00	19	0.09
Gln	787	0.92	122	0.71	222	1.30	159	0.93	50	0.29	234	1.37
Arg	878	0.74	179	0.75	319	1.34	189	0.79	79	0.33	112	0.47
Ser	1,162	0.86	174	0.64	204	0.75	230	0.85	80	0.29	474	1.75
Thr	983	0.78	229	0.91	228	0.91	125	0.50	28	0.11	373	1.48
Val	1,323	0.80	489	1.47	281	0.84	237	0.71	4	0.01	312	0.94
Trp	264	0.80	96	1.46	46	0.70	24	0.36	23	0.35	75	1.14
Tyr	624	0.76	177	1.08	136	0.83	85	0.52	65	0.40	161	0.99
Other	70	0.61	19	0.82	12	0.52	5	0.22	7	0.30	27	1.17

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-5												
Ala	1,172	0.83	312	1.10	300	1.06	259	0.91	159	0.56	142	0.50
Cys	189	0.92	39	0.95	18	0.44	35	0.85	78	1.89	19	0.46
Asp	1,091	1.11	229	1.17	263	1.34	134	0.68	329	1.68	136	0.69
Glu	1,127	1.03	187	0.85	398	1.82	275	1.26	145	0.66	122	0.56
Phe	766	1.15	111	0.83	74	0.55	140	1.05	391	2.92	50	0.37
Gly	461	0.37	144	0.58	98	0.40	43	0.17	55	0.22	121	0.49
His	522	1.33	88	1.12	70	0.89	113	1.44	208	2.66	43	0.55
Ile	666	0.70	135	0.71	94	0.50	211	1.12	177	0.94	49	0.26
Lys	1,152	1.27	228	1.25	346	1.90	307	1.69	97	0.53	174	0.96
Leu	1,213	0.80	340	1.11	230	0.75	316	1.04	259	0.85	68	0.22
Met	225	0.79	59	1.04	52	0.92	56	0.99	47	0.83	11	0.19
Asn	833	1.21	156	1.14	153	1.11	141	1.03	288	2.10	95	0.69
Pro	1,832	2.39	146	0.95	76	0.50	3	0.02	1	0.01	1,606	10.46
Gln	691	1.14	146	1.21	208	1.72	169	1.40	99	0.82	69	0.57
Arg	1,055	1.25	240	1.42	269	1.60	279	1.66	126	0.75	141	0.84
Ser	954	0.99	287	1.49	200	1.04	139	0.72	133	0.69	195	1.02
Thr	794	0.89	176	0.99	192	1.08	211	1.19	93	0.52	122	0.69
Val	827	0.70	133	0.56	127	0.54	258	1.10	234	0.99	75	0.32
Trp	188	0.81	35	0.75	27	0.58	40	0.86	64	1.37	22	0.47
Tyr	657	1.14	90	0.78	82	0.71	155	1.34	300	2.60	30	0.26
Other	65	0.80	15	0.92	19	1.17	12	0.74	13	0.80	6	0.37
type: o5-6												
Ala	1,217	0.78	268	0.85	395	1.26	214	0.68	98	0.31	242	0.77
Cys	226	0.99	67	1.46	28	0.61	54	1.18	28	0.61	49	1.07
Asp	1,340	1.23	398	1.83	199	0.91	420	1.93	144	0.66	179	0.82
Glu	1,167	0.96	298	1.23	410	1.69	179	0.74	77	0.32	203	0.84
Phe	627	0.84	84	0.57	102	0.69	234	1.58	34	0.23	173	1.16
Gly	2,981	2.17	218	0.79	127	0.46	98	0.36	2,248	8.17	290	1.05
His	385	0.89	72	0.83	74	0.85	83	0.95	72	0.83	84	0.97
Ile	608	0.58	151	0.72	137	0.65	74	0.35	2	0.01	244	1.16
Lys	1,202	1.19	277	1.37	305	1.51	158	0.78	200	0.99	262	1.30
Leu	1,414	0.84	281	0.83	316	0.93	396	1.17	62	0.18	359	1.06
Met	222	0.71	48	0.76	42	0.67	54	0.86	22	0.35	56	0.89
Asn	1,010	1.33	242	1.59	129	0.85	202	1.32	285	1.87	152	1.00
Pro	487	0.57	120	0.70	262	1.54	4	0.02	1	0.01	100	0.59
Gln	648	0.97	121	0.90	178	1.33	107	0.80	105	0.78	137	1.02
Arg	869	0.93	187	1.00	223	1.19	111	0.59	144	0.77	204	1.09
Ser	1,254	1.18	293	1.37	220	1.03	446	2.09	73	0.34	222	1.04
Thr	1,111	1.13	218	1.10	166	0.84	515	2.61	11	0.06	201	1.02
Val	736	0.56	169	0.65	185	0.71	73	0.28	11	0.04	298	1.14
Trp	191	0.74	33	0.64	33	0.64	52	1.00	10	0.19	63	1.22
Tyr	522	0.81	99	0.77	109	0.85	169	1.32	24	0.19	121	0.94
Other	78	0.86	15	0.83	19	1.05	16	0.88	8	0.44	20	1.11

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-7												
Ala	957	0.85	251	1.11	194	0.86	181	0.80	154	0.68	177	0.78
Cys	154	0.94	40	1.21	22	0.67	45	1.37	19	0.58	28	0.85
Asp	832	1.06	123	0.79	241	1.54	101	0.65	280	1.79	87	0.56
Glu	760	0.87	114	0.65	190	1.09	207	1.18	151	0.86	98	0.56
Phe	498	0.93	143	1.34	52	0.49	145	1.36	62	0.58	96	0.90
Gly	886	0.90	210	1.06	83	0.42	300	1.52	58	0.29	235	1.19
His	336	1.08	92	1.47	53	0.85	75	1.20	56	0.90	60	0.96
Ile	538	0.71	75	0.50	86	0.57	137	0.91	62	0.41	178	1.18
Lys	716	0.99	79	0.54	211	1.45	185	1.28	138	0.95	103	0.71
Leu	966	0.79	233	0.96	184	0.76	208	0.85	86	0.35	255	1.05
Met	151	0.67	30	0.66	24	0.53	35	0.77	18	0.40	44	0.97
Asn	527	0.96	95	0.87	144	1.31	79	0.72	114	1.04	95	0.87
Pro	1,628	2.66	208	1.70	241	1.97	5	0.04	871	7.11	303	2.47
Gln	472	0.98	81	0.84	108	1.12	116	1.20	77	0.80	90	0.93
Arg	600	0.89	106	0.79	122	0.91	166	1.23	90	0.67	116	0.86
Ser	965	1.26	282	1.84	242	1.58	160	1.04	136	0.89	145	0.95
Thr	802	1.13	176	1.24	225	1.59	160	1.13	101	0.71	140	0.99
Val	756	0.80	126	0.67	103	0.55	182	0.97	81	0.43	264	1.40
Trp	172	0.92	50	1.34	34	0.91	28	0.75	27	0.73	33	0.89
Tyr	407	0.88	112	1.21	65	0.70	110	1.19	42	0.46	78	0.85
Other	32	0.49	5	0.38	7	0.54	6	0.46	8	0.62	6	0.46
type: o5-8												
Ala	992	1.14	152	0.87	47	0.27	186	1.07	221	1.27	386	2.22
Cys	127	1.00	30	1.18	20	0.79	25	0.99	22	0.87	30	1.18
Asp	529	0.88	80	0.66	40	0.33	161	1.34	209	1.73	39	0.32
Glu	624	0.93	89	0.66	23	0.17	212	1.57	208	1.54	92	0.68
Phe	321	0.78	107	1.30	8	0.10	57	0.69	69	0.84	80	0.97
Gly	2,501	3.29	255	1.67	1,648	10.81	91	0.60	235	1.54	272	1.78
His	236	0.98	90	1.87	9	0.19	42	0.87	45	0.93	50	1.04
Ile	375	0.65	125	1.07	3	0.03	109	0.94	51	0.44	87	0.75
Lys	420	0.75	74	0.66	22	0.20	118	1.06	113	1.01	93	0.83
Leu	599	0.64	211	1.13	8	0.04	144	0.77	124	0.66	112	0.60
Met	138	0.79	46	1.32	2	0.06	38	1.09	29	0.83	23	0.66
Asn	339	0.80	53	0.63	21	0.25	70	0.83	96	1.14	99	1.17
Pro	326	0.69	57	0.60	0	0.00	192	2.03	32	0.34	45	0.48
Gln	273	0.73	65	0.87	15	0.20	62	0.83	85	1.14	46	0.62
Arg	361	0.70	80	0.77	21	0.20	83	0.80	71	0.69	106	1.02
Ser	639	1.08	102	0.86	97	0.82	171	1.45	136	1.15	133	1.13
Thr	440	0.81	122	1.12	24	0.22	84	0.77	107	0.98	103	0.94
Val	362	0.50	112	0.77	1	0.01	78	0.54	72	0.50	99	0.68
Trp	149	1.04	44	1.53	4	0.14	35	1.22	26	0.91	40	1.39
Tyr	334	0.94	120	1.69	6	0.08	58	0.82	70	0.99	80	1.13
Other	50	1.00	13	1.30	8	0.80	11	1.10	6	0.60	12	1.20

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-9												
Ala	455	0.48	96	0.50	113	0.59	107	0.56	35	0.18	104	0.54
Cys	116	0.83	46	1.65	30	1.08	20	0.72	5	0.18	15	0.54
Asp	759	1.15	77	0.58	64	0.48	467	3.53	102	0.77	49	0.37
Glu	742	1.00	191	1.29	170	1.15	146	0.99	38	0.26	197	1.33
Phe	416	0.92	149	1.65	134	1.48	68	0.75	19	0.21	46	0.51
Gly	2,146	2.57	92	0.55	55	0.33	279	1.67	1,586	9.47	134	0.80
His	308	1.17	65	1.23	75	1.42	81	1.53	30	0.57	57	1.08
Ile	478	0.75	143	1.12	225	1.76	2	0.02	1	0.01	107	0.84
Lys	806	1.31	103	0.84	161	1.31	112	0.91	31	0.25	399	3.25
Leu	534	0.52	192	0.93	175	0.85	40	0.19	13	0.06	114	0.55
Met	139	0.73	42	1.10	24	0.63	26	0.68	11	0.29	36	0.94
Asn	972	2.10	102	1.10	34	0.37	568	6.12	187	2.02	81	0.87
Pro	145	0.28	61	0.59	73	0.70	0	0.00	0	0.00	11	0.11
Gln	430	1.05	72	0.88	80	0.98	73	0.89	33	0.40	172	2.10
Arg	552	0.97	82	0.72	141	1.24	80	0.70	40	0.35	209	1.84
Ser	468	0.72	121	0.93	100	0.77	75	0.58	63	0.49	109	0.84
Thr	367	0.61	191	1.59	65	0.54	7	0.06	7	0.06	97	0.81
Val	742	0.93	224	1.41	308	1.94	9	0.06	2	0.01	199	1.25
Trp	109	0.69	37	1.17	44	1.40	12	0.38	2	0.06	14	0.44
Tyr	407	1.04	126	1.61	151	1.94	50	0.64	19	0.24	61	0.78
Other	39	0.71	14	1.27	4	0.36	4	0.36	2	0.18	15	1.36
type: o5-10												
Ala	416	0.47	111	0.63	101	0.57	18	0.10	83	0.47	103	0.58
Cys	98	0.76	29	1.13	17	0.66	5	0.19	11	0.43	36	1.40
Asp	734	1.20	51	0.42	456	3.73	79	0.65	70	0.57	78	0.64
Glu	594	0.87	133	0.97	127	0.93	28	0.21	198	1.45	108	0.79
Phe	332	0.80	125	1.50	50	0.60	24	0.29	52	0.62	81	0.97
Gly	2,005	2.60	89	0.58	242	1.57	1,600	10.36	24	0.16	50	0.32
His	284	1.16	81	1.66	78	1.60	18	0.37	59	1.21	48	0.98
Ile	440	0.75	198	1.68	0	0.00	0	0.00	111	0.94	131	1.11
Lys	803	1.42	151	1.33	121	1.07	23	0.20	381	3.36	127	1.12
Leu	557	0.59	155	0.82	35	0.18	12	0.06	119	0.63	236	1.24
Met	97	0.55	22	0.62	15	0.42	2	0.06	31	0.88	27	0.76
Asn	858	2.01	59	0.69	527	6.16	127	1.48	85	0.99	60	0.70
Pro	249	0.52	14	0.15	0	0.00	0	0.00	0	0.00	235	2.46
Gln	391	1.04	66	0.88	80	1.06	19	0.25	169	2.24	57	0.76
Arg	529	1.01	128	1.22	78	0.74	24	0.23	193	1.84	106	1.01
Ser	387	0.65	107	0.89	59	0.49	41	0.34	84	0.70	96	0.80
Thr	327	0.59	84	0.76	9	0.08	4	0.04	95	0.86	135	1.22
Val	681	0.93	282	1.92	5	0.03	1	0.01	191	1.30	202	1.38
Trp	111	0.76	30	1.03	4	0.14	8	0.28	14	0.48	55	1.89
Tyr	351	0.98	135	1.88	47	0.65	20	0.28	71	0.99	78	1.08
Other	26	0.51	4	0.39	3	0.30	1	0.10	13	1.28	5	0.49

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-11												
Ala	414	0.68	116	0.95	79	0.65	37	0.30	93	0.76	89	0.73
Cys	80	0.90	22	1.24	26	1.46	10	0.56	13	0.73	9	0.51
Asp	564	1.34	17	0.20	107	1.27	76	0.90	279	3.30	85	1.01
Glu	442	0.94	48	0.51	109	1.15	41	0.43	137	1.45	107	1.13
Phe	238	0.83	68	1.18	73	1.27	5	0.09	32	0.55	60	1.04
Gly	1,463	2.74	179	1.67	25	0.23	984	9.21	57	0.53	218	2.04
His	177	1.05	34	1.01	64	1.90	19	0.56	19	0.56	41	1.21
Ile	274	0.67	72	0.88	124	1.52	5	0.06	27	0.33	46	0.56
Lys	414	1.06	115	1.47	89	1.14	44	0.56	75	0.96	91	1.16
Leu	383	0.58	103	0.78	110	0.84	8	0.06	62	0.47	100	0.76
Met	87	0.71	25	1.02	16	0.66	6	0.25	18	0.74	22	0.90
Asn	372	1.26	29	0.49	46	0.78	43	0.73	199	3.36	55	0.93
Pro	160	0.48	65	0.98	43	0.65	0	0.00	50	0.76	2	0.03
Gln	245	0.94	64	1.23	53	1.02	13	0.25	44	0.84	71	1.36
Arg	351	0.97	118	1.62	72	0.99	29	0.40	52	0.72	80	1.10
Ser	458	1.11	80	0.97	93	1.12	70	0.85	128	1.55	87	1.05
Thr	328	0.86	87	1.13	57	0.74	1	0.01	76	0.99	107	1.40
Val	336	0.66	103	1.01	146	1.44	3	0.03	17	0.17	67	0.66
Trp	84	0.84	23	1.14	22	1.09	5	0.25	10	0.50	24	1.19
Tyr	202	0.81	46	0.92	63	1.26	11	0.22	30	0.60	52	1.04
Other	33	0.94	7	1.00	4	0.57	11	1.57	3	0.43	8	1.14
type: o5-12												
Ala	414	0.75	100	0.91	129	1.17	98	0.89	2	0.02	85	0.77
Cys	58	0.72	20	1.24	16	1.00	13	0.81	1	0.06	8	0.50
Asp	311	0.82	88	1.15	61	0.80	84	1.10	1	0.01	77	1.01
Glu	381	0.89	65	0.76	136	1.60	100	1.17	1	0.01	79	0.93
Phe	234	0.90	51	0.98	31	0.60	92	1.77	0	0.00	60	1.15
Gly	1,547	3.21	136	1.41	41	0.42	48	0.50	1,251	12.97	71	0.74
His	124	0.81	24	0.79	15	0.49	51	1.67	0	0.00	34	1.12
Ile	209	0.57	35	0.48	42	0.57	49	0.67	0	0.00	83	1.13
Lys	327	0.92	67	0.95	103	1.46	84	1.19	2	0.03	71	1.00
Leu	445	0.75	87	0.73	100	0.84	146	1.23	1	0.01	111	0.94
Met	68	0.62	17	0.77	13	0.59	19	0.86	0	0.00	19	0.86
Asn	252	0.94	82	1.53	41	0.77	87	1.63	3	0.06	39	0.73
Pro	359	1.20	60	1.00	174	2.91	5	0.08	0	0.00	120	2.01
Gln	188	0.80	33	0.70	50	1.06	63	1.34	1	0.02	41	0.87
Arg	266	0.81	56	0.85	81	1.23	66	1.01	2	0.03	61	0.93
Ser	362	0.97	123	1.65	97	1.30	64	0.86	8	0.11	70	0.94
Thr	329	0.95	143	2.07	52	0.75	47	0.68	5	0.07	82	1.18
Val	230	0.50	33	0.36	64	0.70	34	0.37	0	0.00	99	1.08
Trp	70	0.77	19	1.05	14	0.77	18	0.99	0	0.00	19	1.05
Tyr	216	0.96	39	0.87	18	0.40	105	2.33	4	0.09	50	1.11
Other	25	0.79	5	0.79	5	0.79	10	1.58	1	0.16	4	0.63

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-13												
Ala	352	0.67	97	0.93	104	0.99	21	0.20	63	0.60	67	0.64
Cys	57	0.75	22	1.44	5	0.33	5	0.33	15	0.98	10	0.66
Asp	314	0.87	35	0.48	79	1.09	78	1.08	36	0.50	86	1.19
Glu	349	0.86	60	0.74	124	1.53	15	0.19	74	0.91	76	0.94
Phe	195	0.79	67	1.35	36	0.73	16	0.32	39	0.79	37	0.75
Gly	1,110	2.43	95	1.04	23	0.25	787	8.59	127	1.39	78	0.85
His	152	1.05	31	1.07	19	0.66	28	0.97	46	1.59	28	0.97
Ile	210	0.60	59	0.84	56	0.80	1	0.01	39	0.56	55	0.79
Lys	448	1.34	62	0.92	115	1.71	18	0.27	175	2.61	78	1.16
Leu	334	0.59	144	1.28	68	0.60	7	0.06	46	0.41	69	0.61
Met	79	0.76	24	1.15	10	0.48	7	0.33	26	1.24	12	0.57
Asn	301	1.19	32	0.63	41	0.81	132	2.60	45	0.89	51	1.00
Pro	411	1.45	78	1.38	230	4.06	0	0.00	0	0.00	103	1.82
Gln	183	0.82	23	0.51	45	1.01	17	0.38	66	1.48	32	0.72
Arg	278	0.89	62	1.00	49	0.79	25	0.40	84	1.35	58	0.93
Ser	331	0.93	58	0.82	60	0.85	30	0.42	78	1.10	105	1.48
Thr	312	0.95	77	1.17	36	0.55	9	0.14	99	1.51	91	1.39
Val	371	0.85	121	1.39	65	0.75	1	0.01	79	0.91	105	1.21
Trp	63	0.73	20	1.16	14	0.81	3	0.17	11	0.64	15	0.87
Tyr	215	1.01	44	1.03	37	0.87	17	0.40	60	1.41	57	1.34
Other	25	0.83	7	1.16	2	0.33	1	0.17	10	1.66	5	0.83
type: o5-14												
Ala	440	0.95	94	1.02	133	1.44	43	0.47	85	0.92	85	0.92
Cys	61	0.91	17	1.26	14	1.04	12	0.89	5	0.37	13	0.97
Asp	448	1.40	60	0.94	54	0.84	111	1.74	144	2.25	79	1.24
Glu	330	0.93	45	0.63	73	1.02	27	0.38	88	1.23	97	1.36
Phe	180	0.83	43	0.99	36	0.82	16	0.37	25	0.57	60	1.37
Gly	576	1.43	174	2.15	46	0.57	224	2.77	56	0.69	76	0.94
His	148	1.16	25	0.98	42	1.64	24	0.94	17	0.67	40	1.57
Ile	161	0.52	31	0.50	43	0.70	12	0.19	22	0.36	53	0.86
Lys	247	0.83	40	0.67	75	1.27	20	0.34	67	1.13	45	0.76
Leu	359	0.72	111	1.12	80	0.80	27	0.27	48	0.48	93	0.94
Met	59	0.64	13	0.70	15	0.81	5	0.27	14	0.76	12	0.65
Asn	250	1.12	42	0.94	44	0.98	61	1.36	43	0.96	60	1.34
Pro	288	1.15	67	1.34	26	0.52	4	0.08	182	3.64	9	0.18
Gln	151	0.77	28	0.71	34	0.86	22	0.56	24	0.61	43	1.09
Arg	228	0.83	50	0.91	71	1.29	17	0.31	44	0.80	46	0.84
Ser	573	1.83	87	1.39	90	1.44	220	3.51	100	1.60	76	1.21
Thr	454	1.57	56	0.97	86	1.48	201	3.47	42	0.72	69	1.19
Val	200	0.52	34	0.44	60	0.78	8	0.10	37	0.48	61	0.79
Trp	57	0.75	18	1.18	13	0.85	7	0.46	10	0.66	9	0.59
Tyr	140	0.74	33	0.88	33	0.88	14	0.37	17	0.45	43	1.14
Other	25	0.94	7	1.32	7	1.32	0	0.00	5	0.94	6	1.13

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-15												
Ala	379	0.92	58	0.70	113	1.37	103	1.25	41	0.50	64	0.78
Cys	67	1.12	19	1.58	17	1.41	11	0.92	6	0.50	14	1.16
Asp	344	1.21	48	0.84	56	0.98	85	1.49	118	2.07	37	0.65
Glu	276	0.87	43	0.67	29	0.45	92	1.44	66	1.03	46	0.72
Phe	147	0.75	32	0.82	25	0.64	15	0.38	55	1.41	20	0.51
Gly	318	0.88	86	1.19	138	1.91	30	0.42	10	0.14	54	0.75
His	141	1.24	33	1.45	27	1.18	28	1.23	35	1.53	18	0.79
Ile	162	0.59	35	0.64	6	0.11	24	0.44	52	0.94	45	0.82
Lys	263	0.99	59	1.11	36	0.68	80	1.51	52	0.98	36	0.68
Leu	279	0.63	84	0.95	41	0.46	62	0.70	42	0.47	50	0.56
Met	56	0.68	14	0.85	6	0.36	12	0.73	11	0.67	13	0.79
Asn	254	1.27	42	1.05	33	0.82	30	0.75	122	3.05	27	0.67
Pro	587	2.63	64	1.43	124	2.77	117	2.62	1	0.02	281	6.29
Gln	144	0.82	37	1.05	18	0.51	22	0.62	40	1.13	27	0.77
Arg	230	0.94	50	1.02	37	0.75	51	1.04	57	1.16	35	0.71
Ser	363	1.30	69	1.23	105	1.88	69	1.23	58	1.04	62	1.11
Thr	291	1.12	51	0.98	81	1.56	45	0.87	66	1.27	48	0.93
Val	280	0.82	76	1.11	25	0.36	55	0.80	79	1.15	45	0.66
Trp	59	0.87	23	1.69	12	0.88	9	0.66	7	0.52	8	0.59
Tyr	147	0.87	32	0.95	31	0.92	17	0.51	40	1.19	27	0.80
Other	13	0.55	5	1.05	0	0.00	3	0.63	2	0.42	3	0.63
type: o5-16												
Ala	234	0.52	59	0.65	20	0.22	48	0.53	43	0.47	64	0.70
Cys	60	0.91	20	1.51	1	0.08	9	0.68	13	0.98	17	1.28
Asp	478	1.52	79	1.26	62	0.99	245	3.90	71	1.13	21	0.33
Glu	325	0.93	79	1.12	35	0.50	96	1.37	80	1.14	35	0.50
Phe	173	0.81	58	1.35	4	0.09	11	0.26	53	1.23	47	1.10
Gly	880	2.22	26	0.33	749	9.42	41	0.52	29	0.36	35	0.44
His	120	0.96	40	1.59	17	0.68	14	0.56	39	1.55	10	0.40
Ile	214	0.71	77	1.27	3	0.05	8	0.13	49	0.81	77	1.27
Lys	302	1.04	64	1.10	39	0.67	63	1.08	80	1.37	56	0.96
Leu	328	0.67	77	0.79	10	0.10	27	0.28	96	0.98	118	1.21
Met	56	0.62	14	0.77	3	0.17	8	0.44	17	0.94	14	0.77
Asn	333	1.51	48	1.09	40	0.91	170	3.86	44	1.00	31	0.70
Pro	114	0.46	11	0.22	0	0.00	44	0.89	0	0.00	59	1.20
Gln	197	1.02	50	1.29	18	0.46	31	0.80	60	1.55	38	0.98
Arg	272	1.01	56	1.04	23	0.43	51	0.94	81	1.50	61	1.13
Ser	304	0.99	69	1.12	17	0.28	98	1.59	67	1.09	53	0.86
Thr	264	0.93	37	0.65	0	0.00	54	0.95	91	1.60	82	1.44
Val	368	0.98	131	1.73	3	0.04	11	0.15	70	0.93	153	2.03
Trp	66	0.88	14	0.94	3	0.20	7	0.47	19	1.27	23	1.54
Tyr	182	0.98	46	1.24	8	0.22	19	0.51	50	1.35	59	1.59
Other	15	0.57	2	0.38	2	0.38	2	0.38	5	0.96	4	0.77

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-17												
Ala	292	0.80	92	1.26	57	0.78	53	0.72	24	0.33	66	0.90
Cys	57	1.07	9	0.84	20	1.88	5	0.47	8	0.75	15	1.41
Asp	294	1.16	79	1.56	29	0.57	90	1.78	76	1.50	20	0.40
Glu	204	0.72	48	0.85	27	0.48	84	1.49	21	0.37	24	0.42
Phe	183	1.06	35	1.01	63	1.82	39	1.13	23	0.67	23	0.67
Gly	598	1.87	62	0.97	24	0.38	23	0.36	418	6.53	71	1.11
His	106	1.05	20	0.99	18	0.89	20	0.99	29	1.43	19	0.94
Ile	159	0.65	35	0.72	54	1.11	45	0.92	1	0.02	24	0.49
Lys	237	1.01	54	1.15	31	0.66	73	1.56	16	0.34	63	1.34
Leu	360	0.92	58	0.74	161	2.04	53	0.67	23	0.29	65	0.83
Met	57	0.78	11	0.75	17	1.16	11	0.75	6	0.41	12	0.82
Asn	273	1.54	35	0.99	30	0.85	28	0.79	119	3.36	61	1.72
Pro	125	0.63	57	1.44	29	0.73	9	0.23	0	0.00	30	0.76
Gln	159	1.02	31	0.99	25	0.80	47	1.50	16	0.51	40	1.28
Arg	231	1.06	45	1.03	30	0.69	53	1.22	27	0.62	76	1.75
Ser	234	0.94	54	1.09	48	0.97	53	1.07	16	0.32	63	1.27
Thr	233	1.02	45	0.98	69	1.50	48	1.05	1	0.02	70	1.52
Val	230	0.76	40	0.66	52	0.85	61	1.00	1	0.02	76	1.25
Trp	55	0.91	17	1.41	18	1.49	9	0.75	6	0.50	5	0.42
Tyr	151	1.01	21	0.70	40	1.34	46	1.54	20	0.67	24	0.80
Other	17	0.81	3	0.71	9	2.14	1	0.24	0	0.00	4	0.95
type: o5-18												
Ala	220	0.74	38	0.64	67	1.13	8	0.14	60	1.02	47	0.80
Cys	46	1.07	15	1.74	3	0.35	4	0.46	14	1.63	10	1.16
Asp	146	0.71	30	0.73	35	0.86	17	0.42	22	0.54	42	1.03
Glu	218	0.95	28	0.61	64	1.40	5	0.11	49	1.07	72	1.58
Phe	143	1.02	45	1.61	14	0.50	9	0.32	31	1.11	44	1.58
Gly	717	2.77	50	0.97	17	0.33	532	10.28	30	0.58	88	1.70
His	71	0.87	12	0.73	15	0.92	7	0.43	21	1.28	16	0.98
Ile	108	0.55	37	0.94	23	0.58	0	0.00	18	0.46	30	0.76
Lys	203	1.07	46	1.21	67	1.77	10	0.26	65	1.71	15	0.40
Leu	233	0.73	82	1.29	28	0.44	9	0.14	64	1.01	50	0.79
Met	59	1.00	12	1.01	8	0.68	4	0.34	21	1.78	14	1.18
Asn	137	0.96	20	0.70	18	0.63	33	1.15	37	1.29	29	1.01
Pro	234	1.46	55	1.72	156	4.87	0	0.00	0	0.00	23	0.72
Gln	121	0.96	21	0.83	34	1.35	4	0.16	40	1.58	22	0.87
Arg	150	0.85	36	1.02	34	0.97	11	0.31	50	1.42	19	0.54
Ser	167	0.83	28	0.70	35	0.87	13	0.32	54	1.35	37	0.92
Thr	132	0.71	31	0.84	15	0.40	2	0.05	50	1.35	34	0.92
Val	160	0.65	52	1.06	26	0.53	0	0.00	26	0.53	56	1.14
Trp	34	0.70	9	0.92	7	0.72	2	0.21	8	0.82	8	0.82
Tyr	130	1.08	38	1.58	22	0.91	17	0.70	23	0.95	30	1.24
Other	11	0.65	3	0.88	0	0.00	1	0.29	5	1.47	2	0.59

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-19												
Ala	187	0.96	34	0.87	39	1.00	37	0.95	54	1.39	23	0.59
Cys	23	0.81	2	0.35	1	0.18	5	0.88	11	1.94	4	0.71
Asp	186	1.38	28	1.04	30	1.11	52	1.93	24	0.89	52	1.93
Glu	147	0.98	16	0.53	42	1.40	36	1.20	22	0.73	31	1.03
Phe	92	1.00	16	0.87	9	0.49	20	1.09	33	1.79	14	0.76
Gly	169	0.99	87	2.55	15	0.44	8	0.23	28	0.82	31	0.91
His	48	0.89	6	0.56	11	1.02	15	1.39	10	0.93	6	0.56
Ile	85	0.65	9	0.35	10	0.38	35	1.35	9	0.35	22	0.85
Lys	142	1.14	12	0.48	49	1.96	36	1.44	21	0.84	24	0.96
Leu	140	0.67	41	0.98	32	0.76	43	1.03	11	0.26	13	0.31
Met	37	0.95	3	0.39	7	0.90	12	1.54	10	1.28	5	0.64
Asn	135	1.43	44	2.33	19	1.01	21	1.11	21	1.11	30	1.59
Pro	99	0.94	24	1.14	21	1.00	0	0.00	0	0.00	54	2.56
Gln	72	0.87	11	0.66	13	0.78	19	1.14	17	1.02	12	0.72
Arg	119	1.03	19	0.82	34	1.47	19	0.82	30	1.30	17	0.73
Ser	209	1.58	37	1.40	42	1.59	26	0.98	56	2.12	48	1.82
Thr	170	1.39	37	1.51	36	1.47	22	0.90	35	1.43	40	1.64
Val	85	0.53	8	0.25	26	0.80	18	0.56	17	0.53	16	0.49
Trp	26	0.81	3	0.47	5	0.78	8	1.25	8	1.25	2	0.31
Tyr	88	1.11	15	0.94	11	0.69	20	1.26	33	2.08	9	0.57
Other	6	0.54	1	0.45	1	0.45	1	0.45	3	1.34	0	0.00
type: o5-20												
Ala	157	0.88	48	1.35	45	1.27	2	0.06	24	0.67	38	1.07
Cys	12	0.46	4	0.77	2	0.39	1	0.19	1	0.19	4	0.77
Asp	145	1.18	39	1.58	68	2.76	4	0.16	20	0.81	14	0.57
Glu	133	0.97	43	1.56	26	0.95	5	0.18	25	0.91	34	1.24
Phe	54	0.64	12	0.71	3	0.18	4	0.24	21	1.25	14	0.83
Gly	534	3.43	34	1.09	13	0.42	372	11.95	10	0.32	105	3.37
His	33	0.67	9	0.92	6	0.61	5	0.51	10	1.02	3	0.31
Ile	71	0.60	16	0.67	10	0.42	0	0.00	24	1.01	21	0.88
Lys	104	0.91	29	1.27	23	1.01	7	0.31	33	1.45	12	0.53
Leu	152	0.79	29	0.76	32	0.84	1	0.03	53	1.38	37	0.97
Met	35	0.98	4	0.56	13	1.83	1	0.14	9	1.26	8	1.12
Asn	101	1.17	19	1.10	35	2.03	5	0.29	30	1.74	12	0.70
Pro	48	0.50	11	0.57	3	0.16	0	0.00	0	0.00	34	1.76
Gln	67	0.88	17	1.12	18	1.18	1	0.07	21	1.38	10	0.66
Arg	95	0.90	25	1.18	22	1.04	3	0.14	32	1.51	13	0.61
Ser	108	0.90	30	1.24	39	1.62	2	0.08	25	1.04	12	0.50
Thr	82	0.73	12	0.54	29	1.30	1	0.04	28	1.25	12	0.54
Val	67	0.45	15	0.51	12	0.41	0	0.00	21	0.71	19	0.64
Trp	14	0.48	3	0.51	4	0.68	0	0.00	3	0.51	4	0.68
Tyr	48	0.66	15	1.03	8	0.55	0	0.00	18	1.24	7	0.48
Other	10	0.98	0	0.00	3	1.47	0	0.00	6	2.93	1	0.49

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-21												
Ala	126	0.77	33	1.01	26	0.79	14	0.43	32	0.98	21	0.64
Cys	24	1.01	9	1.89	6	1.26	2	0.42	2	0.42	5	1.05
Asp	134	1.19	67	2.96	37	1.63	8	0.35	7	0.31	15	0.66
Glu	72	0.57	11	0.43	18	0.71	4	0.16	16	0.63	23	0.91
Phe	59	0.76	16	1.03	14	0.90	5	0.32	9	0.58	15	0.97
Gly	438	3.06	35	1.22	83	2.90	264	9.21	17	0.59	39	1.36
His	54	1.20	10	1.10	19	2.10	6	0.66	14	1.55	5	0.55
Ile	42	0.38	6	0.27	1	0.05	0	0.00	10	0.46	25	1.14
Lys	104	0.99	7	0.33	25	1.19	3	0.14	64	3.05	5	0.24
Leu	105	0.60	20	0.57	13	0.37	1	0.03	28	0.79	43	1.22
Met	30	0.92	3	0.46	4	0.61	7	1.07	9	1.37	7	1.07
Asn	129	1.63	12	0.76	63	3.97	28	1.76	10	0.63	16	1.01
Pro	17	0.19	10	0.56	0	0.00	0	0.00	0	0.00	7	0.39
Gln	71	1.02	6	0.43	13	0.93	6	0.43	33	2.36	13	0.93
Arg	94	0.97	5	0.26	17	0.87	4	0.21	57	2.93	11	0.56
Ser	117	1.06	36	1.62	20	0.90	15	0.68	25	1.13	21	0.95
Thr	134	1.31	68	3.31	3	0.15	3	0.15	19	0.92	41	1.99
Val	57	0.42	9	0.33	4	0.15	0	0.00	12	0.44	32	1.18
Trp	32	1.19	7	1.30	2	0.37	1	0.19	6	1.11	16	2.97
Tyr	56	0.84	9	0.67	11	0.82	8	0.60	10	0.75	18	1.35
Other	10	1.06	2	1.06	2	1.06	2	1.06	1	0.53	3	1.59
type: o5-22												
Ala	87	0.64	25	0.91	20	0.73	1	0.04	15	0.55	26	0.95
Cys	15	0.75	3	0.75	2	0.50	0	0.00	1	0.25	9	2.25
Asp	92	0.97	11	0.58	19	1.00	3	0.16	44	2.32	15	0.79
Glu	81	0.77	8	0.38	36	1.70	0	0.00	31	1.46	6	0.28
Phe	45	0.70	15	1.16	10	0.77	0	0.00	10	0.77	10	0.77
Gly	365	3.05	23	0.96	12	0.50	308	12.84	5	0.21	17	0.71
His	36	0.95	13	1.72	3	0.40	1	0.13	9	1.19	10	1.32
Ile	67	0.73	24	1.31	14	0.76	0	0.00	14	0.76	15	0.82
Lys	73	0.83	13	0.74	29	1.65	0	0.00	16	0.91	15	0.85
Leu	110	0.75	41	1.39	21	0.71	2	0.07	22	0.75	24	0.81
Met	21	0.77	8	1.46	7	1.28	1	0.18	1	0.18	4	0.73
Asn	69	1.04	9	0.68	8	0.60	0	0.00	28	2.11	24	1.81
Pro	88	1.19	17	1.14	20	1.35	0	0.00	19	1.28	32	2.15
Gln	41	0.70	10	0.85	14	1.20	0	0.00	6	0.51	11	0.94
Arg	56	0.69	10	0.61	16	0.98	1	0.06	22	1.35	7	0.43
Ser	71	0.77	21	1.13	22	1.18	0	0.00	13	0.70	15	0.81
Thr	87	1.01	14	0.81	23	1.34	2	0.12	26	1.51	22	1.28
Val	122	1.07	34	1.49	28	1.23	0	0.00	18	0.79	42	1.84
Trp	22	0.98	3	0.66	6	1.33	0	0.00	6	1.33	7	1.55
Tyr	43	0.77	17	1.52	8	0.72	0	0.00	13	1.16	5	0.45
Other	4	0.51	0	0.00	1	0.63	0	0.00	0	0.00	3	1.90

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-23												
Ala	98	0.74	25	0.94	26	0.98	15	0.56	0	0.00	32	1.20
Cys	20	1.03	7	1.80	9	2.32	0	0.00	0	0.00	4	1.03
Asp	47	0.51	10	0.54	11	0.60	17	0.92	2	0.11	7	0.38
Glu	76	0.74	13	0.63	23	1.12	32	1.55	0	0.00	8	0.39
Phe	64	1.02	22	1.75	10	0.79	10	0.79	0	0.00	22	1.75
Gly	51	0.44	14	0.60	13	0.56	1	0.04	1	0.04	22	0.94
His	25	0.68	7	0.95	6	0.81	7	0.95	0	0.00	5	0.68
Ile	54	0.61	23	1.29	6	0.34	13	0.73	0	0.00	12	0.67
Lys	55	0.64	8	0.47	5	0.29	34	1.99	0	0.00	8	0.47
Leu	94	0.66	31	1.08	15	0.52	19	0.66	2	0.07	27	0.94
Met	18	0.68	3	0.56	5	0.94	5	0.94	0	0.00	5	0.94
Asn	56	0.87	12	0.93	6	0.46	22	1.70	2	0.15	14	1.08
Pro	414	5.75	28	1.94	42	2.91	14	0.97	300	20.78	30	2.08
Gln	56	0.99	18	1.58	9	0.79	16	1.41	0	0.00	13	1.14
Arg	48	0.61	8	0.50	14	0.88	11	0.69	1	0.06	14	0.88
Ser	112	1.24	21	1.16	32	1.77	25	1.38	2	0.11	32	1.77
Thr	88	1.05	16	0.96	42	2.51	19	1.14	0	0.00	11	0.66
Val	98	0.89	29	1.31	23	1.04	23	1.04	0	0.00	23	1.04
Trp	15	0.69	1	0.23	5	1.14	7	1.60	0	0.00	2	0.46
Tyr	53	0.98	13	1.20	7	0.64	17	1.56	0	0.00	16	1.47
Other	8	1.05	1	0.65	1	0.65	3	1.96	0	0.00	3	1.96
type: o5-24												
Ala	132	1.07	44	1.78	19	0.77	11	0.45	24	0.97	34	1.38
Cys	17	0.95	2	0.56	6	1.67	2	0.56	2	0.56	5	1.39
Asp	82	0.96	5	0.29	10	0.59	5	0.29	44	2.58	18	1.06
Glu	66	0.69	3	0.16	11	0.58	2	0.10	29	1.52	21	1.10
Phe	41	0.71	11	0.94	11	0.94	1	0.09	6	0.51	12	1.03
Gly	363	3.38	50	2.32	18	0.83	239	11.07	9	0.42	47	2.18
His	34	1.00	3	0.44	12	1.76	2	0.29	5	0.73	12	1.76
Ile	77	0.94	16	0.97	18	1.09	0	0.00	33	2.00	10	0.61
Lys	49	0.62	7	0.44	18	1.14	0	0.00	19	1.20	5	0.32
Leu	81	0.61	46	1.73	8	0.30	1	0.04	16	0.60	10	0.38
Met	20	0.81	4	0.81	1	0.20	0	0.00	8	1.62	7	1.42
Asn	44	0.74	7	0.59	10	0.84	6	0.50	13	1.09	8	0.67
Pro	28	0.42	4	0.30	0	0.00	0	0.00	16	1.20	8	0.60
Gln	46	0.88	6	0.57	21	1.99	1	0.09	8	0.76	10	0.95
Arg	53	0.72	9	0.61	23	1.57	3	0.20	11	0.75	7	0.48
Ser	101	1.21	19	1.14	36	2.15	11	0.66	15	0.90	20	1.20
Thr	82	1.06	21	1.36	35	2.26	0	0.00	8	0.52	18	1.16
Val	57	0.56	13	0.63	14	0.68	0	0.00	13	0.63	17	0.83
Trp	15	0.74	3	0.74	5	1.23	0	0.00	3	0.74	4	0.98
Tyr	37	0.74	12	1.19	10	0.99	3	0.30	3	0.30	9	0.89
Other	10	1.41	2	1.41	1	0.70	0	0.00	2	1.41	5	3.52

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-25												
Ala	80	0.96	17	1.02	13	0.78	30	1.80	0	0.00	20	1.20
Cys	13	1.07	6	2.47	2	0.82	2	0.82	0	0.00	3	1.24
Asp	43	0.75	3	0.26	15	1.30	7	0.61	0	0.00	18	1.56
Glu	39	0.61	7	0.54	8	0.62	13	1.01	0	0.00	11	0.85
Phe	40	1.02	4	0.51	13	1.65	9	1.14	0	0.00	14	1.78
Gly	46	0.63	16	1.10	12	0.82	4	0.27	1	0.07	13	0.89
His	26	1.13	9	1.95	5	1.08	2	0.43	2	0.43	8	1.74
Ile	21	0.38	10	0.90	4	0.36	2	0.18	0	0.00	5	0.45
Lys	51	0.95	5	0.47	9	0.84	28	2.62	0	0.00	9	0.84
Leu	52	0.58	16	0.89	11	0.61	6	0.33	0	0.00	19	1.06
Met	10	0.60	7	2.10	2	0.60	1	0.30	0	0.00	0	0.00
Asn	26	0.64	8	0.99	7	0.87	3	0.37	0	0.00	8	0.99
Pro	263	5.83	10	1.11	23	2.55	41	4.54	189	20.92	0	0.00
Gln	21	0.59	4	0.56	6	0.84	5	0.70	0	0.00	6	0.84
Arg	46	0.93	14	1.41	20	2.02	6	0.60	0	0.00	6	0.60
Ser	40	0.71	9	0.80	13	1.15	6	0.53	0	0.00	12	1.06
Thr	38	0.73	16	1.53	13	1.24	4	0.38	1	0.10	4	0.38
Val	44	0.64	18	1.30	15	1.08	2	0.14	1	0.07	8	0.58
Trp	16	1.17	3	1.09	0	0.00	5	1.82	0	0.00	8	2.91
Tyr	52	1.53	11	1.62	3	0.44	17	2.50	0	0.00	21	3.09
Other	3	0.63	1	1.04	0	0.00	1	1.04	0	0.00	1	1.04
type: o5-26												
Ala	38	0.60	9	0.71	1	0.08	6	0.47	10	0.79	12	0.94
Cys	7	0.76	0	0.00	0	0.00	1	0.54	4	2.16	2	1.08
Asp	63	1.43	40	4.55	3	0.34	5	0.57	8	0.91	7	0.80
Glu	35	0.71	9	0.92	1	0.10	12	1.22	3	0.31	10	1.02
Phe	18	0.60	1	0.17	3	0.50	1	0.17	4	0.67	9	1.50
Gly	134	2.41	3	0.27	120	10.78	3	0.27	2	0.18	6	0.54
His	21	1.19	2	0.57	1	0.28	11	3.13	6	1.71	1	0.28
Ile	26	0.61	0	0.00	0	0.00	7	0.82	10	1.18	9	1.06
Lys	47	1.15	13	1.59	1	0.12	25	3.06	5	0.61	3	0.37
Leu	34	0.50	4	0.29	0	0.00	10	0.73	15	1.10	5	0.37
Met	7	0.55	1	0.39	1	0.39	1	0.39	3	1.18	1	0.39
Asn	45	1.46	12	1.95	8	1.30	5	0.81	13	2.11	7	1.14
Pro	27	0.78	13	1.89	0	0.00	0	0.00	0	0.00	14	2.03
Gln	32	1.18	4	0.74	1	0.18	13	2.39	6	1.10	8	1.47
Arg	27	0.71	7	0.93	1	0.13	9	1.19	5	0.66	5	0.66
Ser	40	0.93	6	0.70	2	0.23	13	1.51	8	0.93	11	1.28
Thr	41	1.03	9	1.13	0	0.00	8	1.00	12	1.50	12	1.50
Val	54	1.02	9	0.85	1	0.09	12	1.13	18	1.70	14	1.32
Trp	12	1.15	2	0.95	0	0.00	1	0.48	3	1.43	6	2.86
Tyr	31	1.20	4	0.77	4	0.77	5	0.96	12	2.31	6	1.16
Other	1	0.27	0	0.00	0	0.00	0	0.00	1	1.37	0	0.00

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-27												
Ala	58	0.86	25	1.84	19	1.40	7	0.52	2	0.15	5	0.37
Cys	7	0.71	2	1.01	0	0.00	2	1.01	0	0.00	3	1.52
Asp	22	0.47	1	0.11	8	0.85	2	0.21	1	0.11	10	1.06
Glu	49	0.94	0	0.00	8	0.76	22	2.10	0	0.00	19	1.81
Phe	15	0.47	6	0.94	3	0.47	6	0.94	0	0.00	0	0.00
Gly	191	3.22	20	1.68	6	0.50	4	0.34	151	12.71	10	0.84
His	14	0.75	4	1.07	1	0.27	6	1.60	0	0.00	3	0.80
Ile	50	1.11	12	1.32	11	1.21	5	0.55	0	0.00	22	2.43
Lys	33	0.76	5	0.57	5	0.57	17	1.95	0	0.00	6	0.69
Leu	57	0.78	16	1.09	21	1.44	10	0.68	1	0.07	9	0.62
Met	7	0.52	1	0.37	3	1.10	1	0.37	0	0.00	2	0.74
Asn	26	0.79	3	0.46	5	0.76	4	0.61	1	0.15	13	1.97
Pro	26	0.71	15	2.04	10	1.36	0	0.00	0	0.00	1	0.14
Gln	27	0.93	3	0.52	4	0.69	12	2.07	0	0.00	8	1.38
Arg	31	0.77	6	0.74	5	0.62	15	1.86	0	0.00	5	0.62
Ser	67	1.46	12	1.30	16	1.74	19	2.06	1	0.11	19	2.06
Thr	35	0.82	2	0.23	15	1.76	8	0.94	0	0.00	10	1.17
Val	46	0.82	20	1.77	15	1.33	6	0.53	0	0.00	5	0.44
Trp	13	1.17	3	1.34	1	0.45	5	2.24	0	0.00	4	1.79
Tyr	11	0.40	0	0.00	1	0.18	7	1.26	1	0.18	2	0.36
Other	5	1.28	2	2.56	1	1.28	0	0.00	0	0.00	2	2.56
type: o5-28												
Ala	46	0.77	18	1.51	0	0.00	9	0.75	4	0.33	15	1.26
Cys	16	1.84	3	1.72	1	0.57	7	4.02	1	0.57	4	2.30
Asp	28	0.68	3	0.36	4	0.48	6	0.73	12	1.45	3	0.36
Glu	31	0.67	5	0.54	5	0.54	10	1.08	6	0.65	5	0.54
Phe	37	1.31	14	2.48	2	0.35	5	0.89	5	0.89	11	1.95
Gly	122	2.34	10	0.96	88	8.42	14	1.34	1	0.10	9	0.86
His	10	0.61	3	0.91	1	0.30	3	0.91	2	0.61	1	0.30
Ile	18	0.45	4	0.50	0	0.00	0	0.00	5	0.63	9	1.13
Lys	43	1.13	9	1.17	4	0.52	22	2.87	5	0.65	3	0.39
Leu	46	0.72	21	1.63	1	0.08	1	0.08	11	0.86	12	0.93
Met	15	1.26	3	1.26	2	0.84	1	0.42	0	0.00	9	3.77
Asn	39	1.35	11	1.90	11	1.90	3	0.52	9	1.55	5	0.86
Pro	45	1.39	3	0.46	0	0.00	0	0.00	30	4.64	12	1.85
Gln	25	0.98	3	0.59	6	1.18	9	1.76	4	0.78	3	0.59
Arg	35	0.99	1	0.14	9	1.27	20	2.81	3	0.42	2	0.28
Ser	33	0.82	6	0.74	3	0.37	11	1.36	10	1.23	3	0.37
Thr	28	0.75	9	1.20	1	0.13	5	0.67	9	1.20	4	0.53
Val	45	0.91	6	0.60	0	0.00	8	0.81	8	0.81	23	2.32
Trp	7	0.71	2	1.02	0	0.00	0	0.00	4	2.03	1	0.51
Tyr	23	0.95	5	1.03	1	0.21	5	1.03	9	1.85	3	0.62
Other	3	0.88	0	0.00	0	0.00	0	0.00	1	1.46	2	2.91

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-29												
Ala	52	0.84	12	0.97	6	0.49	1	0.08	24	1.94	9	0.73
Cys	11	1.22	7	3.88	1	0.55	0	0.00	2	1.11	1	0.55
Asp	20	0.47	4	0.47	12	1.40	0	0.00	1	0.12	3	0.35
Glu	21	0.44	8	0.84	8	0.84	0	0.00	2	0.21	3	0.31
Phe	21	0.72	8	1.37	7	1.20	0	0.00	2	0.34	4	0.68
Gly	24	0.44	10	0.92	0	0.00	0	0.00	4	0.37	10	0.92
His	9	0.53	4	1.17	2	0.58	0	0.00	2	0.58	1	0.29
Ile	31	0.75	6	0.73	8	0.97	0	0.00	6	0.73	11	1.33
Lys	22	0.55	3	0.38	8	1.01	0	0.00	8	1.01	3	0.38
Leu	34	0.51	10	0.75	3	0.23	0	0.00	10	0.75	11	0.83
Met	6	0.48	0	0.00	0	0.00	0	0.00	3	1.21	3	1.21
Asn	31	1.03	2	0.33	12	2.00	1	0.17	6	1.00	10	1.67
Pro	224	6.68	9	1.34	0	0.00	142	21.18	44	6.56	29	4.33
Gln	25	0.95	5	0.95	11	2.08	0	0.00	3	0.57	6	1.13
Arg	18	0.49	3	0.41	3	0.41	0	0.00	3	0.41	9	1.22
Ser	54	1.29	12	1.43	21	2.50	0	0.00	14	1.67	7	0.83
Thr	46	1.18	19	2.45	17	2.19	0	0.00	2	0.26	8	1.03
Val	43	0.84	13	1.26	11	1.07	0	0.00	6	0.58	13	1.26
Trp	5	0.49	1	0.49	2	0.98	0	0.00	1	0.49	1	0.49
Tyr	23	0.91	8	1.59	12	2.38	0	0.00	1	0.20	2	0.40
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
type: o5-30												
Ala	47	0.86	10	0.91	0	0.00	12	1.09	5	0.45	20	1.82
Cys	5	0.62	3	1.87	0	0.00	0	0.00	2	1.25	0	0.00
Asp	17	0.45	3	0.39	1	0.13	1	0.13	11	1.45	1	0.13
Glu	27	0.64	5	0.59	1	0.12	14	1.65	5	0.59	2	0.24
Phe	19	0.73	10	1.92	1	0.19	2	0.38	2	0.38	4	0.77
Gly	221	4.60	17	1.77	125	12.99	1	0.10	57	5.92	21	2.18
His	22	1.45	4	1.32	0	0.00	3	0.99	7	2.30	8	2.63
Ile	21	0.57	10	1.36	0	0.00	9	1.23	0	0.00	2	0.27
Lys	28	0.79	6	0.85	0	0.00	14	1.98	3	0.43	5	0.71
Leu	22	0.37	7	0.59	0	0.00	6	0.51	4	0.34	5	0.42
Met	4	0.36	1	0.45	0	0.00	3	1.36	0	0.00	0	0.00
Asn	45	1.69	9	1.69	0	0.00	5	0.94	18	3.37	13	2.44
Pro	29	0.97	5	0.84	0	0.00	22	3.69	0	0.00	2	0.34
Gln	12	0.51	4	0.85	0	0.00	3	0.64	2	0.43	3	0.64
Arg	23	0.70	7	1.07	0	0.00	7	1.07	3	0.46	6	0.92
Ser	25	0.67	2	0.27	0	0.00	7	0.94	5	0.67	11	1.47
Thr	21	0.61	7	1.01	0	0.00	3	0.43	0	0.00	11	1.59
Val	23	0.50	5	0.55	0	0.00	8	0.87	0	0.00	10	1.09
Trp	5	0.55	1	0.55	0	0.00	2	1.10	1	0.55	1	0.55
Tyr	20	0.89	9	2.01	0	0.00	6	1.34	3	0.67	2	0.45
Other	4	1.27	3	4.74	0	0.00	0	0.00	0	0.00	1	1.58

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-31												
Ala	55	1.04	12	1.14	15	1.42	6	0.57	10	0.95	12	1.14
Cys	6	0.78	2	1.30	2	1.30	1	0.65	0	0.00	1	0.65
Asp	66	1.81	5	0.68	22	3.01	19	2.60	14	1.91	6	0.82
Glu	50	1.22	9	1.10	8	0.98	3	0.37	16	1.96	14	1.71
Phe	17	0.68	3	0.60	5	1.00	0	0.00	5	1.00	4	0.80
Gly	46	0.99	19	2.05	6	0.65	14	1.51	3	0.32	4	0.43
His	24	1.64	2	0.68	8	2.74	6	2.05	1	0.34	7	2.40
Ile	8	0.23	2	0.28	2	0.28	1	0.14	1	0.14	2	0.28
Lys	20	0.59	4	0.59	5	0.74	0	0.00	1	0.15	10	1.47
Leu	32	0.56	10	0.88	4	0.35	0	0.00	5	0.44	13	1.14
Met	7	0.66	2	0.95	3	1.42	0	0.00	1	0.47	1	0.47
Asn	44	1.72	5	0.98	14	2.73	10	1.95	8	1.56	7	1.37
Pro	31	1.08	7	1.22	2	0.35	0	0.00	20	3.49	2	0.35
Gln	13	0.58	1	0.22	5	1.11	2	0.44	2	0.44	3	0.66
Arg	20	0.64	4	0.64	2	0.32	1	0.16	3	0.48	10	1.59
Ser	65	1.81	6	0.84	5	0.70	34	4.74	13	1.81	7	0.98
Thr	42	1.27	8	1.21	2	0.30	18	2.71	8	1.21	6	0.90
Val	25	0.57	9	1.02	3	0.34	3	0.34	7	0.80	3	0.34
Trp	8	0.92	2	1.15	0	0.00	2	1.15	2	1.15	2	1.15
Tyr	34	1.58	10	2.32	9	2.09	3	0.70	3	0.70	9	2.09
Other	2	0.66	1	1.64	1	1.64	0	0.00	0	0.00	0	0.00
type: o5-32												
Ala	33	0.78	5	0.59	9	1.06	1	0.12	10	1.18	8	0.94
Cys	7	1.13	0	0.00	1	0.81	1	0.81	3	2.42	2	1.61
Asp	42	1.43	2	0.34	10	1.70	12	2.04	7	1.19	11	1.87
Glu	30	0.92	6	0.91	6	0.91	6	0.91	7	1.06	5	0.76
Phe	18	0.90	5	1.24	2	0.50	5	1.24	2	0.50	4	1.00
Gly	61	1.65	35	4.70	15	2.01	5	0.67	1	0.13	5	0.67
His	5	0.43	1	0.43	1	0.43	1	0.43	1	0.43	1	0.43
Ile	22	0.78	2	0.35	3	0.53	4	0.70	9	1.58	4	0.70
Lys	27	0.99	3	0.55	8	1.47	6	1.10	3	0.55	7	1.28
Leu	40	0.88	6	0.66	2	0.22	16	1.75	10	1.09	6	0.66
Met	7	0.83	1	0.59	3	1.76	0	0.00	1	0.59	2	1.18
Asn	33	1.61	6	1.45	3	0.73	12	2.91	6	1.45	6	1.45
Pro	19	0.83	7	1.52	0	0.00	0	0.00	0	0.00	12	2.60
Gln	14	0.77	2	0.55	1	0.28	5	1.38	3	0.83	3	0.83
Arg	13	0.52	2	0.40	4	0.79	1	0.20	4	0.79	2	0.40
Ser	22	0.77	3	0.52	2	0.35	8	1.39	4	0.69	5	0.87
Thr	40	1.50	7	1.31	14	2.62	9	1.69	6	1.12	4	0.75
Val	27	0.77	2	0.28	8	1.13	2	0.28	10	1.41	5	0.71
Trp	11	1.58	1	0.71	4	2.86	1	0.71	5	3.57	0	0.00
Tyr	20	1.16	2	0.58	3	0.86	4	1.15	6	1.73	5	1.44
Other	4	1.64	1	2.04	0	0.00	0	0.00	1	2.04	2	4.09

Table A3.9 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 5 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o5-33												
Ala	26	0.56	8	0.85	2	0.21	7	0.75	1	0.11	8	0.85
Cys	5	0.73	4	2.93	0	0.00	1	0.73	0	0.00	0	0.00
Asp	24	0.74	6	0.93	2	0.31	11	1.70	3	0.46	2	0.31
Glu	13	0.36	3	0.41	2	0.28	6	0.83	0	0.00	2	0.28
Phe	8	0.36	1	0.23	3	0.68	0	0.00	1	0.23	3	0.68
Gly	173	4.22	4	0.49	72	8.78	2	0.24	89	10.86	6	0.73
His	11	0.85	2	0.77	3	1.16	2	0.77	0	0.00	4	1.54
Ile	14	0.45	2	0.32	0	0.00	1	0.16	0	0.00	11	1.76
Lys	36	1.20	2	0.33	3	0.50	12	2.00	3	0.50	16	2.66
Leu	32	0.63	5	0.50	3	0.30	7	0.69	0	0.00	17	1.69
Met	6	0.64	3	1.60	0	0.00	3	1.60	0	0.00	0	0.00
Asn	19	0.84	3	0.66	5	1.10	5	1.10	2	0.44	4	0.88
Pro	8	0.32	8	1.58	0	0.00	0	0.00	0	0.00	0	0.00
Gln	30	1.50	14	3.50	2	0.50	10	2.50	4	1.00	0	0.00
Arg	47	1.69	21	3.77	8	1.44	12	2.15	2	0.36	4	0.72
Ser	30	0.94	8	1.26	1	0.16	10	1.57	4	0.63	7	1.10
Thr	31	1.05	8	1.36	0	0.00	13	2.21	0	0.00	10	1.70
Val	18	0.46	3	0.39	1	0.13	7	0.90	0	0.00	7	0.90
Trp	1	0.13	1	0.65	0	0.00	0	0.00	0	0.00	0	0.00
Tyr	13	0.68	3	0.79	2	0.52	0	0.00	0	0.00	8	2.09
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
type: o5-34												
Ala	12	0.28	3	0.35	0	0.00	0	0.00	4	0.47	5	0.58
Cys	1	0.16	1	0.80	0	0.00	0	0.00	0	0.00	0	0.00
Asp	85	2.86	11	1.85	65	10.93	0	0.00	3	0.50	6	1.01
Glu	19	0.57	5	0.75	0	0.00	0	0.00	11	1.66	3	0.45
Phe	16	0.79	11	2.71	1	0.25	1	0.25	1	0.25	2	0.49
Gly	106	2.82	10	1.33	2	0.27	88	11.70	3	0.40	3	0.40
His	20	1.68	6	2.53	0	0.00	0	0.00	7	2.95	7	2.95
Ile	7	0.24	0	0.00	1	0.17	0	0.00	1	0.17	5	0.87
Lys	29	1.05	3	0.54	1	0.18	1	0.18	17	3.08	7	1.27
Leu	12	0.26	2	0.22	0	0.00	3	0.32	5	0.54	2	0.22
Met	6	0.70	1	0.58	0	0.00	3	1.75	0	0.00	2	1.16
Asn	48	2.30	6	1.44	23	5.52	1	0.24	3	0.72	15	3.60
Pro	3	0.13	2	0.43	1	0.21	0	0.00	0	0.00	0	0.00
Gln	15	0.82	1	0.27	0	0.00	0	0.00	9	2.45	5	1.36
Arg	13	0.51	1	0.20	0	0.00	1	0.20	6	1.17	5	0.98
Ser	24	0.82	13	2.23	0	0.00	0	0.00	9	1.54	2	0.34
Thr	45	1.67	9	1.67	6	1.11	1	0.19	10	1.85	19	3.52
Val	19	0.53	2	0.28	0	0.00	0	0.00	9	1.26	8	1.12
Trp	9	1.27	8	5.65	0	0.00	0	0.00	0	0.00	1	0.71
Tyr	11	0.63	5	1.43	0	0.00	1	0.29	2	0.57	3	0.86
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

OPEN 6 TURNS

Table A3.10. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. Turn types o6-23 and o6-48 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		ψ_{i+1}		φ_{i+1}		ω_{i+1}		ψ_{i+2}		φ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o6-23	571	179.0	5.3	76.4	16.0	8.3	23.5	179.2	5.4	-107.5	29.4	153.3	23.4
o6-23	*398*	*178.8*	*6.1*	*76.6*	*15.8*	*8.4*	*22.1*	*179.3*	*5.7*	*-107.1*	*29.1*	*154.6*	*23.8*
o6-24	478	-179.7	4.8	-65.0	17.0	130.2	14.4	178.7	4.1	74.1	14.7	9.6	20.0
o6-24	*320*	*-179.7*	*5.0*	*-65.2*	*17.3*	*129.3*	*14.6*	*178.7*	*3.9*	*75.2*	*14.8*	*9.1*	*20.9*
o6-25	526	-174.9	5.7	61.8	9.8	-124.3	8.9	179.9	3.6	-94.0	12.5	5.1	13.5
o6-25	*369*	*-175.0*	*5.8*	*61.8*	*10.2*	*-124.0*	*9.7*	*179.9*	*4.2*	*-93.0*	*12.7*	*2.8*	*13.9*
o6-26	475	177.7	5.4	-77.3	21.6	145.7	17.6	179.6	5.7	-55.5	7.0	133.1	9.3
o6-26	*311*	*178.0*	*5.4*	*-78.3*	*22.6*	*145.6*	*17.5*	*179.3*	*6.2*	*-55.4*	*7.2*	*132.4*	*8.5*
o6-27	433	177.3	5.6	-114.1	28.7	171.7	18.3	-179.9	4.4	-61.0	6.8	-26.7	10.5
o6-27	*300*	*177.0*	*6.2*	*-114.7*	*28.8*	*171.8*	*19.2*	*180.0*	*4.5*	*-61.7*	*8.1*	*-26.1*	*11.3*
o6-28	312	175.5	5.8	-72.9	13.6	160.6	14.4	-176.4	5.6	-59.6	10.9	-30.1	11.9
o6-28	*198*	*175.3*	*5.4*	*-71.2*	*14.1*	*159.2*	*14.5*	*-177.0*	*5.2*	*-59.2*	*10.1*	*-29.3*	*11.0*
o6-29	256	-179.4	3.6	-66.9	10.2	-33.8	11.9	-177.9	3.8	-83.7	12.4	-36.9	11.8
o6-29	*181*	*-179.4*	*4.0*	*-67.5*	*10.3*	*-33.0*	*12.6*	*-177.9*	*4.1*	*-84.9*	*12.4*	*-36.0*	*12.1*
o6-30	258	179.0	5.0	-58.2	9.4	134.6	9.8	178.7	3.9	78.5	12.9	5.8	16.9
o6-30	*201*	*179.1*	*4.8*	*-58.2*	*9.7*	*134.1*	*10.2*	*178.5*	*3.3*	*78.8*	*12.9*	*5.4*	*15.5*
o6-31	158	-177.6	6.1	-99.9	19.4	6.2	20.3	177.7	5.3	70.8	22.0	-147.3	22.1
o6-31	*116*	*-178.1*	*5.4*	*-99.4*	*19.2*	*4.7*	*22.0*	*178.1*	*4.9*	*69.2*	*22.2*	*-146.4*	*21.2*
o6-32	131	-179.9	4.4	-92.0	21.7	148.5	18.7	174.9	6.2	-81.8	23.5	145.1	9.7
o6-32	*88*	*179.5*	*4.9*	*-93.0*	*22.8*	*146.7*	*18.6*	*175.2*	*6.1*	*-85.3*	*24.4*	*144.7*	*10.3*
o6-33	130	178.4	6.7	-101.3	20.4	173.4	18.5	177.9	5.6	-59.7	11.1	147.9	13.8
o6-33	*83*	*178.7*	*6.6*	*-102.4*	*21.9*	*174.6*	*19.2*	*177.9*	*6.3*	*-59.0*	*12.0*	*146.6*	*13.7*
o6-34	101	178.6	4.5	-83.7	19.7	-11.2	15.2	-179.4	4.3	-115.6	15.1	35.0	11.9
o6-34	*57*	*180.0*	*3.7*	*-78.9*	*15.8*	*-14.0*	*13.2*	*179.9*	*4.7*	*-112.3*	*16.1*	*33.4*	*13.2*
o6-35	115	177.1	5.8	-98.9	20.6	118.8	12.4	179.2	6.1	-129.3	14.4	118.3	12.3
o6-35	*73*	*176.5*	*6.5*	*-96.7*	*19.6*	*118.0*	*14.8*	*178.8*	*5.3*	*-127.4*	*15.6*	*119.1*	*11.6*
o6-36	88	-178.2	4.7	-100.2	18.0	-7.4	15.9	179.7	5.4	-110.5	32.6	154.3	17.3
o6-36	*64*	*-177.7*	*4.9*	*-96.5*	*15.9*	*-9.1*	*14.7*	*179.5*	*5.6*	*-109.7*	*33.9*	*155.0*	*15.8*
o6-37	113	180.0	6.1	-110.5	19.3	118.5	14.0	-179.0	4.3	-56.3	5.6	132.1	5.5
o6-37	*80*	*-179.5*	*5.6*	*-108.0*	*20.7*	*120.0*	*13.8*	*-179.4*	*4.8*	*-56.7*	*5.5*	*133.1*	*5.3*
o6-38	105	179.9	5.8	-97.3	20.0	99.9	29.4	-179.4	4.8	-161.8	32.9	163.0	17.6
o6-38	*68*	*179.5*	*5.5*	*-100.7*	*21.0*	*103.8*	*30.1*	*-179.9*	*5.1*	*-166.5*	*37.4*	*165.0*	*17.3*
o6-39	94	-179.1	4.5	-64.1	13.1	138.8	19.8	177.2	5.7	60.5	9.8	26.0	14.6
o6-39	*56*	*-178.7*	*5.6*	*-66.3*	*14.2*	*143.3*	*21.9*	*176.9*	*6.7*	*58.7*	*7.9*	*27.2*	*15.4*
o6-40	87	-179.7	3.4	84.2	12.1	-4.3	11.1	179.1	4.8	-117.8	13.5	-61.7	13.4
o6-40	*66*	*-178.7*	*3.7*	*83.5*	*13.1*	*-4.7*	*11.1*	*178.8*	*5.0*	*-115.3*	*14.5*	*-61.7*	*13.3*
o6-41	81	-179.0	5.4	-104.6	12.9	9.8	10.3	-177.9	4.2	98.1	12.7	-32.7	25.8
o6-41	*58*	*-179.2*	*4.9*	*-104.2*	*13.1*	*9.7*	*10.8*	*-177.3*	*4.0*	*97.1*	*11.3*	*-28.4*	*23.1*
o6-42	85	179.1	6.8	76.6	13.3	-166.5	22.2	-176.8	4.4	-63.1	7.3	-25.0	11.1
o6-42	*57*	*-179.5*	*6.9*	*74.1*	*13.8*	*-162.0*	*22.0*	*-176.3*	*4.9*	*-64.4*	*7.2*	*-25.0*	*11.6*
o6-43	85	-180.0	4.8	-69.1	16.4	142.5	13.4	178.5	3.7	72.9	12.2	15.4	14.3
o6-43	*63*	*-179.6*	*4.5*	*-70.4*	*21.7*	*142.2*	*14.2*	*177.5*	*4.4*	*73.1*	*13.8*	*15.8*	*16.0*
o6-44	73	-177.8	5.7	-94.0	19.6	-8.9	16.8	-178.8	4.0	-60.6	9.7	143.3	14.0
o6-44	*52*	*-177.1*	*6.7*	*-96.1*	*18.9*	*-6.3*	*18.0*	*-179.3*	*4.6*	*-61.8*	*9.7*	*144.3*	*12.9*
o6-45	70	179.7	5.7	120.8	33.6	-165.2	14.6	179.0	4.4	-75.8	11.9	174.3	16.3
o6-45	*56*	*179.6*	*5.2*	*120.9*	*34.0*	*-166.1*	*15.8*	*178.0*	*5.1*	*-74.2*	*12.4*	*173.0*	*15.4*
o6-46	54	-179.5	3.7	-72.0	10.6	-34.3	11.6	-177.6	4.9	-91.6	14.8	-19.3	16.0
o6-46	*44*	*-179.3*	*3.4*	*-73.5*	*11.0*	*-35.3*	*10.9*	*-177.0*	*4.2*	*-90.5*	*15.8*	*-22.3*	*15.9*
o6-47	61	178.5	4.4	55.5	7.2	27.5	9.4	-179.3	4.0	66.0	10.0	13.9	12.4
o6-47	*39*	*178.4*	*4.6*	*54.9*	*6.8*	*29.0*	*7.8*	*-178.6*	*3.9*	*64.3*	*8.5*	*15.3*	*10.7*
o6-48	60	177.8	5.6	101.4	19.2	3.6	14.3	176.0	4.7	74.1	23.1	-156.3	21.9
o6-48	*49*	*178.3*	*6.2*	*98.3*	*18.2*	*5.8*	*15.8*	*176.3*	*4.7*	*75.0*	*26.0*	*-155.1*	*24.2*

Table A3.10 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. Turn types o6-23 and o6-48 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+2}		ψ_{i+3}		φ_{i+3}		ω_{i+3}		ψ_{i+4}		φ_{i+4}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o6-23	571	-179.2	5.2	-60.9	7.8	-32.8	12.2	-179.6	4.0	-71.7	15.1	-25.8	20.6
o6-23	*398*	*-179.4*	*5.5*	*-60.9*	*7.9*	*-32.2*	*12.5*	*-180.0*	*3.8*	*-70.9*	*14.3*	*-26.9*	*19.8*
o6-24	478	179.9	5.2	-126.7	21.6	159.6	17.8	175.7	6.1	-85.4	21.7	132.6	22.2
o6-24	*320*	*-179.9*	*6.5*	*-128.2*	*22.4*	*159.3*	*18.3*	*175.6*	*6.1*	*-84.2*	*21.5*	*134.0*	*21.1*
o6-25	526	-177.7	5.1	-105.4	19.4	143.3	19.4	175.1	5.5	-93.4	21.5	132.6	14.4
o6-25	*369*	*-177.8*	*5.5*	*-103.2*	*20.9*	*143.5*	*20.0*	*175.1*	*5.4*	*-94.1*	*22.3*	*132.3*	*14.4*
o6-26	475	179.4	3.8	83.0	13.4	-8.9	19.8	-178.1	4.7	-96.3	25.7	-8.5	30.8
o6-26	*311*	*179.8*	*3.5*	*83.7*	*13.8*	*-8.1*	*18.7*	*-178.2*	*4.8*	*-97.7*	*26.2*	*-7.4*	*32.3*
o6-27	433	-179.0	4.1	-94.0	14.7	-0.1	16.4	179.9	4.7	124.9	29.6	-172.3	23.9
o6-27	*300*	*-179.0*	*3.9*	*-94.1*	*14.0*	*-0.6*	*15.6*	*-179.7*	*4.8*	*123.3*	*28.8*	*-171.7*	*24.5*
o6-28	312	-179.1	4.0	-91.9	12.9	2.7	14.3	-178.6	5.7	-105.0	20.1	155.9	18.5
o6-28	*198*	*-179.1*	*3.8*	*-92.2*	*13.0*	*2.3*	*14.7*	*-178.5*	*5.7*	*-104.3*	*20.2*	*155.4*	*18.7*
o6-29	256	-172.1	5.5	-117.6	12.9	-7.6	15.3	-177.7	4.8	-106.7	30.4	172.8	24.0
o6-29	*181*	*-172.5*	*5.1*	*-117.4*	*12.6*	*-5.8*	*13.9*	*-177.8*	*4.9*	*-111.1*	*29.1*	*175.5*	*25.3*
o6-30	258	178.7	4.4	-106.5	22.8	164.8	12.7	179.3	5.5	-72.8	17.2	-22.5	16.3
o6-30	*201*	*178.7*	*4.6*	*-105.5*	*22.0*	*166.5*	*14.0*	*179.1*	*5.2*	*-72.8*	*18.7*	*-22.2*	*17.7*
o6-31	158	-177.2	4.5	-62.8	8.9	-24.8	14.5	179.3	3.8	-80.9	17.9	-13.2	21.2
o6-31	*116*	*-177.2*	*3.8*	*-62.8*	*10.4*	*-25.5*	*13.9*	*179.7*	*4.2*	*-78.9*	*16.5*	*-16.4*	*21.2*
o6-32	131	3.2	4.4	-89.6	7.5	-2.3	12.4	-175.7	4.8	-65.6	12.9	-32.8	13.1
o6-32	*88*	*3.1*	*4.7*	*-89.2*	*7.2*	*-4.1*	*11.7*	*-175.9*	*4.8*	*-64.1*	*11.9*	*-33.8*	*13.0*
o6-33	130	179.0	5.7	69.4	14.9	22.4	16.7	176.6	5.5	67.7	13.6	26.5	16.1
o6-33	*83*	*179.2*	*5.3*	*69.9*	*15.3*	*20.9*	*17.8*	*177.6*	*4.9*	*66.8*	*11.9*	*28.9*	*14.5*
o6-34	101	176.0	3.8	65.0	13.1	31.6	11.4	176.8	4.2	82.9	16.8	11.3	17.9
o6-34	*57*	*176.1*	*3.8*	*66.6*	*14.4*	*31.7*	*11.8*	*177.5*	*4.0*	*80.1*	*17.1*	*13.3*	*19.3*
o6-35	115	-178.0	5.2	51.7	4.7	41.1	7.9	177.2	3.3	81.6	9.6	1.5	10.7
o6-35	*73*	*-178.7*	*4.9*	*52.4*	*5.3*	*41.0*	*6.4*	*176.5*	*3.2*	*82.6*	*8.7*	*0.1*	*9.8*
o6-36	88	176.6	5.2	-118.6	26.2	136.7	14.0	-2.4	4.4	-73.4	8.0	155.5	14.2
o6-36	*64*	*176.8*	*4.9*	*-115.2*	*28.0*	*137.9*	*14.2*	*-1.5*	*4.4*	*-75.0*	*8.1*	*156.1*	*14.2*
o6-37	113	-179.8	3.1	82.3	12.2	-4.9	12.4	-179.5	4.8	-119.9	11.3	-58.8	13.6
o6-37	*80*	*-179.6*	*3.2*	*81.4*	*10.7*	*-5.3*	*12.8*	*-179.6*	*4.2*	*-119.1*	*12.3*	*-59.7*	*13.1*
o6-38	105	179.0	4.8	-57.3	8.4	131.4	7.7	178.6	4.4	85.9	13.4	-3.3	16.9
o6-38	*68*	*178.9*	*5.1*	*-59.4*	*11.3*	*132.7*	*9.2*	*179.3*	*4.6*	*84.4*	*15.5*	*-1.4*	*16.9*
o6-39	94	179.9	5.2	59.4	11.1	29.0	13.5	-176.9	4.5	-70.7	15.2	-24.1	14.6
o6-39	*56*	*179.7*	*5.1*	*60.3*	*11.2*	*27.9*	*13.8*	*-177.1*	*4.6*	*-68.7*	*14.6*	*-24.4*	*13.1*
o6-40	87	178.4	4.1	-91.0	16.1	-27.9	15.9	179.6	4.9	-138.5	20.3	153.8	13.9
o6-40	*66*	*178.5*	*4.3*	*-92.7*	*15.9*	*-29.0*	*18.0*	*179.6*	*4.4*	*-138.1*	*20.5*	*153.8*	*16.2*
o6-41	81	179.4	3.1	-110.1	23.4	-31.2	17.4	-178.8	4.9	-138.6	17.0	156.5	15.0
o6-41	*58*	*179.1*	*4.4*	*-112.5*	*21.0*	*-35.2*	*15.4*	*-179.0*	*4.6*	*-136.5*	*17.9*	*157.6*	*13.7*
o6-42	85	-178.8	4.4	-98.2	10.7	9.3	12.7	179.4	4.9	-75.3	32.1	147.5	22.4
o6-42	*57*	*-178.1*	*5.3*	*-99.0*	*10.6*	*7.0*	*12.1*	*179.0*	*4.3*	*-73.8*	*28.9*	*148.1*	*21.4*
o6-43	85	-178.8	3.8	-105.4	18.8	-0.4	13.9	-179.6	4.6	74.4	13.5	17.8	15.0
o6-43	*63*	*-179.8*	*4.3*	*-106.8*	*22.8*	*-2.6*	*14.6*	*-179.2*	*4.6*	*74.4*	*16.6*	*19.2*	*16.7*
o6-44	73	178.4	5.3	59.8	8.6	30.5	13.7	179.0	4.5	65.4	14.6	27.0	16.9
o6-44	*52*	*178.0*	*4.9*	*59.1*	*7.8*	*33.4*	*12.7*	*179.4*	*3.1*	*64.8*	*14.3*	*26.7*	*19.7*
o6-45	70	-178.5	5.4	-56.4	5.4	-37.3	8.1	179.7	3.0	-65.7	10.0	-34.8	14.5
o6-45	*56*	*-178.4*	*5.3*	*-57.3*	*5.0*	*-35.5*	*8.2*	*179.4*	*3.6*	*-67.1*	*11.8*	*-33.1*	*16.5*
o6-46	54	-177.8	3.8	-118.6	12.3	-86.9	31.1	-179.8	4.3	-177.0	28.1	169.6	19.0
o6-46	*44*	*-177.5*	*4.4*	*-117.3*	*10.9*	*-83.9*	*30.3*	*179.8*	*4.5*	*179.0*	*28.2*	*168.5*	*18.4*
o6-47	61	-175.3	5.2	-121.1	21.0	-27.0	15.8	-177.3	6.3	-131.5	22.9	144.6	21.9
o6-47	*39*	*-174.8*	*5.1*	*-121.0*	*21.0*	*-29.9*	*14.2*	*-177.6*	*5.1*	*-132.4*	*24.1*	*146.1*	*21.1*
o6-48	60	-177.8	4.1	-61.9	8.2	-32.7	13.0	179.2	3.9	-63.0	8.5	-39.4	13.8
o6-48	*49*	*-177.0*	*3.9*	*-62.1*	*6.5*	*-32.1*	*10.6*	*179.6*	*3.3*	*-61.9*	*8.2*	*-39.3*	*12.7*

Table A3.10 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. Turn types o6-23 and o6-48 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+4}		Cα-Cα[Å]	
		∅	±	∅	±
o6-23	571	179.0	4.1	6.5	0.9
o6-23	*398*	*179.1*	*4.1*	*6.5*	*0.9*
o6-24	478	179.6	5.9	7.2	0.7
o6-24	*320*	*179.1*	*5.6*	*7.1*	*0.7*
o6-25	526	179.6	5.3	7.1	0.6
o6-25	*369*	*179.5*	*5.6*	*7.1*	*0.6*
o6-26	475	-179.0	5.0	6.6	1.0
o6-26	*311*	*-179.1*	*5.1*	*6.6*	*1.0*
o6-27	433	-178.7	4.3	5.1	1.1
o6-27	*300*	*-178.6*	*4.8*	*5.2*	*1.1*
o6-28	312	177.7	6.7	7.3	0.6
o6-28	*198*	*177.5*	*6.0*	*7.3*	*0.5*
o6-29	256	178.6	6.8	7.4	0.7
o6-29	*181*	*178.2*	*6.0*	*7.4*	*0.7*
o6-30	258	-179.1	4.7	7.1	0.8
o6-30	*201*	*-179.6*	*4.3*	*7.0*	*0.8*
o6-31	158	-179.5	5.6	7.2	0.7
o6-31	*116*	*-179.3*	*4.9*	*7.3*	*0.7*
o6-32	131	179.7	4.3	5.5	0.7
o6-32	*88*	*-179.4*	*3.8*	*5.6*	*0.7*
o6-33	130	179.0	5.0	6.3	0.8
o6-33	*83*	*178.4*	*5.0*	*6.3*	*0.8*
o6-34	101	178.6	5.0	7.4	0.5
o6-34	*57*	*179.4*	*5.2*	*7.3*	*0.5*
o6-35	115	-179.2	4.4	7.8	0.2
o6-35	*73*	*-179.2*	*4.1*	*7.8*	*0.2*
o6-36	88	178.8	6.3	6.7	0.9
o6-36	*64*	*179.6*	*6.6*	*6.7*	*0.8*
o6-37	113	-179.2	4.9	5.8	0.5
o6-37	*80*	*-179.1*	*4.7*	*5.8*	*0.5*
o6-38	105	-179.8	4.3	6.6	1.0
o6-38	*68*	*179.9*	*4.9*	*6.6*	*1.0*
o6-39	94	179.0	4.8	5.9	0.7
o6-39	*56*	*179.0*	*4.1*	*6.0*	*0.8*
o6-40	87	177.2	7.4	6.1	0.7
o6-40	*66*	*176.6*	*7.5*	*6.1*	*0.7*
o6-41	81	174.7	6.8	4.7	0.9
o6-41	*58*	*174.2*	*6.8*	*4.6*	*0.8*
o6-42	85	179.4	5.4	7.2	0.7
o6-42	*57*	*179.8*	*5.7*	*7.3*	*0.8*
o6-43	85	-179.7	4.1	7.2	0.6
o6-43	*63*	*180.0*	*4.1*	*7.1*	*0.7*
o6-44	73	-179.8	3.9	6.4	1.0
o6-44	*52*	*178.6*	*4.1*	*6.2*	*0.9*
o6-45	70	179.1	4.2	7.2	0.5
o6-45	*56*	*179.2*	*3.8*	*7.3*	*0.6*
o6-46	54	179.5	4.0	5.7	1.1
o6-46	*44*	*-179.9*	*4.1*	*5.5*	*1.1*
o6-47	61	177.0	5.8	6.4	0.7
o6-47	*39*	*177.3*	*5.5*	*6.3*	*0.7*
o6-48	60	179.0	3.9	6.8	0.8
o6-48	*49*	*179.7*	*3.5*	*6.6*	*0.8*

Table A3.11. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. Turn types o6-49 and o6-74 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		ψ_{i+1}		φ_{i+1}		ω_{i+1}		ψ_{i+2}		φ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o6-49	69	179.1	4.5	-105.1	20.3	146.5	21.7	178.2	4.5	55.6	7.9	-128.7	7.9
o6-49	*47*	*179.3*	*4.2*	*-107.1*	*20.0*	*148.6*	*22.3*	*178.8*	*5.2*	*56.0*	*8.5*	*-127.8*	*8.4*
o6-50	61	178.0	5.4	-98.1	21.4	123.6	11.4	178.7	5.3	55.8	9.6	26.8	10.7
o6-50	*34*	*179.5*	*5.7*	*-97.3*	*22.4*	*121.6*	*13.5*	*179.0*	*4.6*	*54.0*	*5.0*	*29.5*	*7.0*
o6-51	56	-177.7	5.4	-90.3	13.5	-17.5	13.8	-178.6	5.0	-144.7	13.8	171.7	11.2
o6-51	*46*	*-175.8*	*5.0*	*-89.1*	*12.9*	*-17.8*	*13.1*	*-177.8*	*4.8*	*-148.3*	*14.3*	*170.3*	*9.0*
o6-52	53	176.0	5.7	-71.9	13.4	138.0	11.7	-179.3	5.2	-75.5	12.0	-17.9	11.7
o6-52	*37*	*177.1*	*4.5*	*-72.5*	*13.6*	*136.8*	*10.4*	*-178.5*	*4.0*	*-76.8*	*13.3*	*-16.0*	*13.6*
o6-53	59	178.3	4.4	-151.7	23.7	148.7	17.9	179.6	4.0	-114.4	10.6	134.4	10.2
o6-53	*39*	*178.2*	*5.5*	*-153.9*	*24.4*	*149.3*	*19.8*	*179.0*	*3.7*	*-113.4*	*12.1*	*134.7*	*10.4*
o6-54	54	176.5	5.8	-92.7	14.3	-172.9	19.1	-179.8	4.6	-73.2	10.2	-19.3	16.6
o6-54	*34*	*177.9*	*5.6*	*-94.6*	*16.6*	*-170.1*	*22.6*	*179.8*	*4.7*	*-72.5*	*10.4*	*-19.5*	*14.3*
o6-55	50	179.5	4.0	81.6	15.4	3.2	16.5	177.4	5.3	-89.3	16.6	148.2	18.6
o6-55	*30*	*179.9*	*6.1*	*79.7*	*18.1*	*6.2*	*18.5*	*177.1*	*5.4*	*-93.5*	*17.6*	*142.6*	*19.1*
o6-56	51	-177.0	3.9	-53.8	11.2	127.4	5.7	-179.9	2.2	76.2	12.7	1.5	14.1
o6-56	*40*	*-176.8*	*4.5*	*-52.7*	*7.4*	*128.7*	*5.2*	*-179.7*	*2.1*	*76.0*	*11.6*	*0.2*	*11.9*
o6-57	47	-179.7	5.7	-85.7	8.0	68.9	10.7	-177.3	5.3	-149.5	15.3	162.2	12.0
o6-57	*37*	*-179.3*	*5.6*	*-86.8*	*6.6*	*69.1*	*12.6*	*-177.0*	*5.0*	*-149.6*	*12.3*	*163.0*	*10.4*
o6-58	36	179.3	3.8	-69.9	11.6	-22.1	13.0	179.4	4.1	-94.5	14.2	6.3	11.8
o6-58	*29*	*179.6*	*4.5*	*-69.3*	*12.0*	*-23.3*	*12.8*	*178.7*	*3.4*	*-90.3*	*13.3*	*3.2*	*10.5*
o6-59	40	178.7	5.1	84.2	20.2	8.1	16.8	177.3	4.8	-97.7	22.4	168.0	12.0
o6-59	*31*	*178.8*	*5.0*	*82.7*	*22.1*	*12.4*	*20.9*	*178.5*	*4.6*	*-102.4*	*22.9*	*172.3*	*12.8*
o6-60	37	-178.5	5.8	-103.5	14.6	8.1	16.4	-177.7	5.3	-94.4	19.4	151.4	18.1
o6-60	*23*	*-178.1*	*4.0*	*-100.7*	*11.2*	*3.1*	*13.3*	*-177.4*	*3.9*	*-92.5*	*18.0*	*156.2*	*20.5*
o6-61	44	-179.9	3.3	-58.9	6.5	131.4	6.1	-179.1	3.0	87.8	11.9	-10.1	12.1
o6-61	*35*	*-179.6*	*2.7*	*-59.3*	*5.9*	*132.6*	*6.9*	*-178.6*	*3.0*	*83.9*	*11.7*	*-9.1*	*14.2*
o6-62	34	175.3	6.3	-116.5	17.2	120.7	8.6	178.4	4.8	-138.4	11.3	110.4	13.0
o6-62	*21*	*178.2*	*5.5*	*-118.0*	*12.6*	*119.8*	*10.0*	*178.4*	*4.2*	*-134.5*	*8.5*	*106.7*	*11.9*
o6-63	42	179.8	5.9	-76.4	26.1	148.7	10.2	-179.9	3.7	-69.3	11.3	-31.7	12.8
o6-63	*29*	*178.6*	*6.0*	*-75.4*	*24.5*	*147.9*	*9.9*	*179.7*	*3.5*	*-70.4*	*13.0*	*-30.0*	*13.4*
o6-64	34	175.9	5.3	-137.0	12.4	-126.2	12.4	173.5	4.1	-65.5	9.5	-33.5	9.7
o6-64	*22*	*175.5*	*7.9*	*-138.6*	*14.2*	*-123.0*	*15.7*	*173.5*	*5.4*	*-64.5*	*8.5*	*-36.6*	*10.0*
o6-65	45	168.8	9.5	-70.0	11.1	151.5	13.6	179.9	4.5	-83.0	16.7	-55.4	18.4
o6-65	*33*	*169.0*	*8.6*	*-70.6*	*10.9*	*151.8*	*14.0*	*179.7*	*4.5*	*-84.0*	*17.8*	*-59.4*	*17.1*
o6-66	34	179.8	5.1	-93.0	13.5	-8.1	12.0	-179.8	4.7	-97.7	22.3	151.5	12.7
o6-66	*26*	*-179.7*	*4.0*	*-92.0*	*11.4*	*-6.2*	*12.3*	*-179.4*	*4.6*	*-103.4*	*26.7*	*150.4*	*11.6*
o6-67	43	179.6	5.2	48.0	6.7	56.7	8.2	-174.1	5.2	-57.9	4.7	-34.0	7.3
o6-67	*39*	*179.7*	*5.7*	*50.1*	*7.6*	*55.2*	*9.2*	*-173.2*	*5.0*	*-57.8*	*5.5*	*-36.2*	*8.8*
o6-68	27	176.6	7.0	-70.6	13.5	148.8	6.2	1.9	2.9	-88.0	5.5	-2.0	9.7
o6-68	*17*	*177.9*	*8.1*	*-74.7*	*16.4*	*149.5*	*6.1*	*2.8*	*3.6*	*-86.4*	*7.8*	*-4.7*	*11.2*
o6-69	23	178.2	4.9	57.6	6.6	30.4	10.7	178.3	2.9	76.5	10.7	8.2	10.5
o6-69	*19*	*178.6*	*4.3*	*56.8*	*6.2*	*31.1*	*10.6*	*179.8*	*3.6*	*73.6*	*11.6*	*9.9*	*12.3*
o6-70	30	179.9	4.9	83.7	8.1	172.8	12.4	-175.5	4.9	-64.2	7.8	-26.9	9.2
o6-70	*21*	*178.6*	*4.7*	*81.5*	*9.0*	*174.6*	*11.2*	*-174.9*	*5.1*	*-63.6*	*8.0*	*-26.8*	*10.9*
o6-71	33	-1.7	6.3	-77.0	12.2	160.6	29.4	-174.5	6.1	-57.6	5.7	-37.7	8.2
o6-71	*31*	*-1.9*	*7.4*	*-78.3*	*10.3*	*164.4*	*30.5*	*-173.6*	*6.5*	*-58.3*	*5.9*	*-37.4*	*7.5*
o6-72	28	-177.4	4.8	97.0	14.0	-16.0	11.6	179.6	2.9	-120.5	15.6	-34.9	13.4
o6-72	*27*	*-178.2*	*5.4*	*93.6*	*11.4*	*-14.6*	*12.6*	*179.9*	*3.0*	*-114.8*	*16.0*	*-35.2*	*12.8*
o6-73	27	-178.4	4.4	-77.7	11.0	157.3	17.5	178.5	5.4	-65.5	12.3	159.6	17.7
o6-73	*18*	*-179.1*	*2.9*	*-76.8*	*16.3*	*159.7*	*17.0*	*-179.3*	*3.5*	*-66.0*	*10.3*	*156.7*	*14.0*
o6-74	30	178.5	5.5	-83.0	15.0	133.1	29.1	176.2	3.7	-54.9	5.4	143.0	5.4
o6-74	*21*	*179.3*	*6.6*	*-87.2*	*18.6*	*131.3*	*31.6*	*176.8*	*3.2*	*-53.5*	*5.2*	*141.5*	*5.8*

Table A3.11 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. Turn types o6-49 and o6-74 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+2}		ψ_{i+3}		φ_{i+3}		ω_{i+3}		ψ_{i+4}		φ_{i+4}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o6-49	69	-178.8	3.2	-95.4	9.9	10.4	12.7	-179.6	4.6	-69.4	10.2	145.4	14.7
o6-49	*47*	*-179.6*	*3.5*	*-96.7*	*10.9*	*9.0*	*12.9*	*179.4*	*4.7*	*-69.8*	*9.7*	*143.7*	*15.9*
o6-50	61	-179.2	4.2	63.1	9.8	16.1	13.2	-174.0	5.1	-128.2	15.4	-32.8	24.1
o6-50	*34*	*-179.1*	*4.6*	*62.0*	*7.5*	*16.2*	*11.6*	*-173.1*	*5.2*	*-125.5*	*15.7*	*-34.3*	*22.7*
o6-51	56	174.9	6.1	74.2	14.7	-160.6	20.4	-176.7	4.2	-60.1	8.0	-34.2	10.2
o6-51	*46*	*174.3*	*5.4*	*70.1*	*13.5*	*-154.0*	*18.8*	*-176.1*	*3.9*	*-62.1*	*6.2*	*-33.7*	*10.3*
o6-52	53	-176.5	5.3	-108.9	17.6	-15.7	18.5	180.0	5.4	-141.7	12.8	147.3	19.6
o6-52	*37*	*-178.0*	*4.3*	*-107.4*	*19.6*	*-13.5*	*19.4*	*179.5*	*5.6*	*-144.2*	*14.1*	*149.1*	*22.5*
o6-53	59	176.1	4.3	59.2	6.9	-119.8	7.2	175.7	3.9	-54.4	5.2	-35.4	5.6
o6-53	*39*	*176.2*	*3.8*	*58.4*	*6.0*	*-120.2*	*6.4*	*175.8*	*3.6*	*-53.9*	*4.6*	*-35.9*	*5.8*
o6-54	54	176.0	5.5	-99.4	18.4	134.7	17.5	-179.4	4.8	-57.6	7.8	138.4	10.6
o6-54	*34*	*175.8*	*5.4*	*-104.1*	*18.2*	*137.0*	*20.5*	*-179.9*	*5.2*	*-58.1*	*8.9*	*138.6*	*13.7*
o6-55	50	179.7	5.3	-57.4	6.4	131.7	10.4	179.9	4.8	81.3	16.9	1.4	16.5
o6-55	*30*	*179.2*	*6.3*	*-56.4*	*6.7*	*134.3*	*7.5*	*179.3*	*3.6*	*79.9*	*13.6*	*0.0*	*18.6*
o6-56	51	-177.9	4.6	-121.1	12.8	-49.8	13.1	-178.7	5.0	-151.9	12.7	148.6	14.1
o6-56	*40*	*-178.5*	*3.3*	*-118.7*	*13.9*	*-50.9*	*14.2*	*-179.6*	*4.9*	*-150.0*	*15.1*	*149.8*	*16.2*
o6-57	47	-179.0	4.5	-69.7	10.0	-30.8	11.4	179.5	3.6	-140.0	14.7	136.5	26.8
o6-57	*37*	*-179.5*	*4.7*	*-70.2*	*9.2*	*-31.5*	*12.0*	*178.7*	*3.5*	*-136.5*	*12.2*	*136.9*	*25.3*
o6-58	36	178.0	3.7	111.3	13.0	0.0	11.2	178.5	3.5	94.4	16.1	174.6	22.4
o6-58	*29*	*179.2*	*4.3*	*109.8*	*13.4*	*1.7*	*10.8*	*177.4*	*3.3*	*98.3*	*14.9*	*171.7*	*20.8*
o6-59	40	-179.3	4.0	-65.3	9.9	-31.7	13.2	175.8	5.1	-127.1	21.2	147.6	19.5
o6-59	*31*	*-179.5*	*4.8*	*-65.8*	*8.6*	*-32.8*	*9.8*	*175.5*	*4.1*	*-130.6*	*20.4*	*149.6*	*15.2*
o6-60	37	176.1	5.8	53.8	11.2	-129.0	9.5	-178.4	4.7	-85.6	13.1	3.2	16.1
o6-60	*23*	*176.3*	*4.9*	*55.9*	*14.1*	*-126.7*	*10.5*	*179.7*	*3.0*	*-84.9*	*14.0*	*1.9*	*14.4*
o6-61	44	179.6	3.6	-115.4	11.0	-59.5	9.1	-179.4	3.5	-98.8	20.4	-11.8	17.7
o6-61	*35*	*179.0*	*4.8*	*-114.0*	*11.5*	*-59.3*	*8.3*	*-179.3*	*3.1*	*-98.5*	*18.6*	*-13.3*	*17.4*
o6-62	34	-175.1	4.3	57.4	6.1	-124.2	6.0	179.3	2.8	-89.9	10.0	2.7	13.7
o6-62	*21*	*-175.5*	*4.6*	*58.7*	*5.4*	*-123.6*	*6.6*	*179.2*	*2.3*	*-89.6*	*9.2*	*2.4*	*13.4*
o6-63	42	179.2	4.2	-132.1	17.5	152.9	16.1	177.9	4.0	56.6	11.3	37.7	17.7
o6-63	*29*	*-179.7*	*4.6*	*-135.1*	*19.7*	*151.1*	*15.0*	*178.5*	*4.0*	*61.1*	*14.2*	*31.8*	*24.1*
o6-64	34	174.3	5.4	-122.2	10.7	171.7	11.5	176.9	6.8	-111.6	21.7	131.4	13.8
o6-64	*22*	*175.2*	*6.7*	*-121.1*	*11.1*	*175.7*	*13.8*	*178.2*	*6.9*	*-116.5*	*23.2*	*134.3*	*17.4*
o6-65	45	-176.9	5.6	-68.3	8.9	120.6	10.7	179.9	2.9	79.8	9.9	2.2	11.3
o6-65	*33*	*-176.4*	*4.3*	*-66.5*	*7.2*	*121.2*	*10.8*	*179.8*	*3.7*	*82.2*	*9.0*	*0.6*	*9.3*
o6-66	34	174.5	6.8	-70.4	21.0	145.4	10.3	2.2	3.4	-86.9	7.5	-0.6	14.0
o6-66	*26*	*175.1*	*6.6*	*-67.5*	*19.2*	*145.4*	*7.5*	*2.6*	*4.3*	*-85.9*	*6.6*	*-2.0*	*12.7*
o6-67	43	179.6	2.8	-66.4	7.7	-35.9	12.7	178.8	3.1	-83.4	22.9	-31.0	16.4
o6-67	*39*	*-179.2*	*3.2*	*-66.0*	*7.7*	*-36.1*	*11.9*	*179.3*	*3.6*	*-82.7*	*21.2*	*-32.8*	*14.5*
o6-68	27	-177.9	3.6	-62.1	6.7	-41.1	7.4	178.9	6.0	-128.0	14.0	145.6	20.0
o6-68	*17*	*-178.6*	*5.3*	*-63.2*	*8.3*	*-39.9*	*7.8*	*-180.0*	*5.2*	*-129.3*	*15.2*	*148.7*	*23.6*
o6-69	23	179.4	3.7	-89.0	19.5	155.4	12.7	179.7	5.3	-77.3	18.0	-16.7	20.5
o6-69	*19*	*178.9*	*3.7*	*-85.9*	*15.7*	*154.5*	*11.8*	*-178.8*	*4.9*	*-74.5*	*16.7*	*-22.9*	*21.2*
o6-70	30	-179.5	5.0	-112.5	14.1	9.3	12.0	178.6	4.6	66.7	11.9	21.3	13.2
o6-70	*21*	*-179.1*	*5.6*	*-112.3*	*16.5*	*8.8*	*13.9*	*178.1*	*4.9*	*69.4*	*14.6*	*19.1*	*16.6*
o6-71	33	-178.0	3.2	-62.2	5.2	-32.9	11.3	178.8	2.6	-74.9	11.2	-34.7	13.9
o6-71	*31*	*-178.5*	*2.6*	*-62.7*	*5.8*	*-32.9*	*12.0*	*179.7*	*2.6*	*-76.0*	*14.3*	*-33.4*	*15.2*
o6-72	28	179.7	3.7	-146.8	12.5	167.2	14.3	176.4	6.2	-82.8	20.7	136.5	9.5
o6-72	*27*	*-179.5*	*4.9*	*-146.1*	*18.2*	*165.9*	*13.1*	*175.7*	*6.5*	*-78.2*	*19.5*	*136.7*	*9.2*
o6-73	27	179.0	4.5	73.4	15.0	-136.7	6.3	-177.8	5.1	-68.6	11.9	-20.9	15.6
o6-73	*18*	*-179.9*	*4.1*	*67.8*	*11.8*	*-136.8*	*7.2*	*-178.5*	*5.6*	*-72.4*	*14.4*	*-15.6*	*18.3*
o6-74	30	8.4	5.8	-85.1	6.2	6.3	6.0	177.5	4.8	-70.1	7.7	151.6	9.7
o6-74	*21*	*9.0*	*4.4*	*-85.3*	*5.0*	*5.1*	*5.0*	*177.2*	*4.4*	*-69.6*	*9.4*	*149.2*	*9.4*

Table A3.11 *(continued)*. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. Turn types o6-49 and o6-74 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_{i+4}		Cα-Cα [Å]	
		∅	±	∅	±
o6-49	69	178.2	4.6	6.5	0.8
o6-49	*47*	*179.2*	*4.4*	*6.3*	*0.8*
o6-50	61	-176.7	5.7	5.5	0.4
o6-50	*34*	*-176.3*	*5.5*	*5.5*	*0.5*
o6-51	56	-179.1	3.0	5.6	1.1
o6-51	*46*	*-178.8*	*4.5*	*5.3*	*1.1*
o6-52	53	-179.7	5.8	7.3	0.6
o6-52	*37*	*-179.6*	*5.1*	*7.3*	*0.7*
o6-53	59	177.7	3.0	7.6	0.3
o6-53	*39*	*177.4*	*2.7*	*7.6*	*0.3*
o6-54	54	178.2	4.2	7.4	0.6
o6-54	*34*	*177.1*	*4.9*	*7.2*	*0.8*
o6-55	50	179.5	5.8	6.8	0.8
o6-55	*30*	*179.4*	*5.4*	*7.0*	*0.7*
o6-56	51	175.6	5.9	6.9	0.6
o6-56	*40*	*175.7*	*6.0*	*6.9*	*0.7*
o6-57	47	178.8	5.1	6.7	0.9
o6-57	*37*	*177.6*	*5.5*	*6.7*	*0.9*
o6-58	36	-178.9	4.5	7.2	0.6
o6-58	*29*	*-178.1*	*5.3*	*7.2*	*0.6*
o6-59	40	178.5	7.2	7.0	0.8
o6-59	*31*	*177.6*	*5.2*	*6.9*	*0.8*
o6-60	37	179.7	5.5	5.9	1.0
o6-60	*23*	*-179.9*	*4.4*	*6.0*	*1.0*
o6-61	44	178.2	5.2	6.7	0.5
o6-61	*35*	*-180.0*	*4.8*	*6.7*	*0.5*
o6-62	34	-178.4	5.5	7.8	0.2
o6-62	*21*	*-179.0*	*3.6*	*7.9*	*0.2*
o6-63	42	-179.4	5.4	6.6	0.7
o6-63	*29*	*-179.3*	*5.0*	*6.7*	*0.7*
o6-64	34	-180.0	5.2	6.4	0.6
o6-64	*22*	*-179.7*	*5.0*	*6.4*	*0.6*
o6-65	45	179.7	5.0	7.1	0.6
o6-65	*33*	*-179.2*	*5.1*	*7.2*	*0.5*
o6-66	34	-178.3	6.5	5.8	1.1
o6-66	*26*	*-178.5*	*6.3*	*5.8*	*1.1*
o6-67	43	-178.8	4.7	7.6	0.4
o6-67	*39*	*-178.2*	*3.9*	*7.6*	*0.3*
o6-68	27	177.9	5.3	7.0	0.8
o6-68	*17*	*178.5*	*5.5*	*6.8*	*0.9*
o6-69	23	179.8	4.8	7.4	0.6
o6-69	*19*	*-179.4*	*4.0*	*7.4*	*0.5*
o6-70	30	-176.5	4.6	6.2	0.5
o6-70	*21*	*-176.1*	*5.5*	*6.1*	*0.6*
o6-71	33	-179.5	5.7	6.7	0.7
o6-71	*31*	*-179.7*	*4.8*	*6.6*	*0.7*
o6-72	28	177.7	5.3	7.0	0.5
o6-72	*27*	*178.4*	*5.6*	*7.1*	*0.5*
o6-73	27	179.4	4.5	6.8	0.8
o6-73	*18*	*179.1*	*4.8*	*6.8*	*0.9*
o6-74	30	-180.0	11.7	6.6	0.8
o6-74	*21*	*178.2*	*5.9*	*6.5*	*0.8*

Table A3.12. The means and standard deviations of the dihedral angles (in degrees) and Cα-Cα distances of the classified turn types for the training and test set of open 6 turn structures. Turn types o6-75 and o6-81 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ω_i		ψ_{i+1}		φ_{i+1}		ω_{i+1}		ψ_{i+2}		φ_{i+2}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o6-75	28	-177.6	6.6	-109.0	16.7	12.7	15.5	-175.5	3.8	-59.8	6.5	-29.1	7.2
o6-75	*15*	*-178.1*	*6.3*	*-106.7*	*17.9*	*14.4*	*17.5*	*-175.4*	*4.3*	*-59.1*	*5.9*	*-29.4*	*7.0*
o6-76	19	178.7	5.2	-105.7	16.3	112.1	17.2	177.2	3.7	-124.5	14.4	-136.5	7.6
o6-76	*11*	*179.2*	*3.7*	*-103.8*	*14.9*	*109.1*	*21.1*	*176.8*	*2.8*	*-116.9*	*15.4*	*-137.6*	*11.0*
o6-77	30	-177.9	4.1	-125.4	5.8	40.2	5.5	176.3	3.7	62.7	8.8	31.3	10.4
o6-77	*13*	*-177.4*	*4.7*	*-125.3*	*5.6*	*41.4*	*4.8*	*177.0*	*2.4*	*62.3*	*5.7*	*33.4*	*9.1*
o6-78	17	177.2	4.7	51.8	15.1	-132.0	10.1	-178.2	3.7	-67.6	8.7	-15.6	9.8
o6-78	*9*	*179.0*	*4.9*	*55.6*	*6.4*	*-133.6*	*8.8*	*-178.3*	*2.6*	*-71.6*	*12.2*	*-12.9*	*9.6*
o6-79	31	-179.4	4.7	56.0	5.5	-127.5	6.2	-179.6	2.2	-83.2	8.7	2.9	8.7
o6-79	*23*	*-178.5*	*3.6*	*56.0*	*4.9*	*-127.6*	*5.7*	*-179.6*	*2.4*	*-82.4*	*8.4*	*2.6*	*11.3*
o6-80	13	-178.2	2.9	-79.0	13.6	-9.1	11.4	177.2	6.7	-92.9	14.6	-38.8	14.9
o6-80	*12*	*-178.2*	*2.4*	*-76.6*	*12.2*	*-12.0*	*10.6*	*176.5*	*5.2*	*-90.3*	*8.8*	*-39.9*	*14.9*
o6-81	13	-178.7	5.0	-94.1	10.2	-3.1	8.6	-179.0	3.4	93.1	7.6	158.9	7.0
o6-81	*14*	*-178.4*	*5.9*	*-93.5*	*12.1*	*-3.8*	*11.8*	*-178.6*	*4.1*	*94.0*	*8.0*	*162.4*	*12.6*

type	|turns|	ω_{i+2}		ψ_{i+3}		φ_{i+3}		ω_{i+3}		ψ_{i+4}		φ_{i+4}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
o6-75	28	179.5	4.0	-101.2	15.8	11.1	12.3	179.4	4.5	88.6	16.1	2.9	16.9
o6-75	*15*	*178.4*	*2.5*	*-100.1*	*18.8*	*8.1*	*11.8*	*-178.5*	*3.9*	*85.7*	*15.8*	*6.0*	*16.4*
o6-76	19	174.4	4.5	-61.6	5.3	-32.0	8.7	174.4	5.4	-126.4	11.5	169.0	8.7
o6-76	*11*	*175.3*	*5.8*	*-59.0*	*8.2*	*-35.6*	*10.4*	*177.9*	*6.4*	*-123.2*	*16.8*	*170.5*	*13.5*
o6-77	30	175.4	4.5	90.1	7.2	3.8	6.7	177.9	3.2	-71.6	7.3	-43.0	7.6
o6-77	*13*	*175.7*	*3.5*	*89.2*	*8.7*	*4.4*	*7.1*	*177.9*	*2.6*	*-72.4*	*5.6*	*-42.0*	*7.4*
o6-78	17	-178.1	8.5	-82.8	7.3	-12.3	13.5	179.8	5.6	82.5	16.2	-162.8	15.4
o6-78	*9*	*179.2*	*4.3*	*-78.1*	*7.4*	*-15.5*	*12.3*	*-177.0*	*5.1*	*80.6*	*12.8*	*-156.0*	*14.4*
o6-79	31	176.1	5.0	87.2	7.6	11.0	9.0	-177.7	3.1	-57.4	5.1	-44.1	10.1
o6-79	*23*	*176.0*	*4.9*	*87.8*	*8.3*	*10.8*	*9.2*	*-177.5*	*3.4*	*-55.9*	*4.5*	*-44.9*	*9.1*
o6-80	13	-178.9	2.5	-87.0	11.2	-30.1	13.1	-178.5	2.9	-159.6	12.2	-175.1	16.0
o6-80	*12*	*-178.5*	*1.6*	*-85.7*	*13.0*	*-27.8*	*10.3*	*-177.8*	*2.3*	*-155.8*	*9.0*	*-171.6*	*17.4*
o6-81	13	177.4	3.1	56.7	5.9	31.5	8.5	178.7	3.2	79.1	8.0	8.1	6.3
o6-81	*14*	*179.4*	*3.2*	*58.7*	*11.0*	*31.1*	*12.7*	*178.1*	*2.9*	*79.2*	*6.4*	*7.5*	*8.7*

type	|turns|	ω_{i+4}		Cα-Cα[Å]	
		∅	±	∅	±
o6-75	28	-179.9	6.3	7.3	0.6
o6-75	*15*	*179.6*	*5.7*	*7.3*	*0.6*
o6-76	19	177.0	5.7	6.1	0.8
o6-76	*11*	*178.8*	*4.6*	*6.3*	*0.9*
o6-77	30	177.7	2.3	7.6	0.1
o6-77	*13*	*178.2*	*2.7*	*7.5*	*0.1*
o6-78	17	-177.1	4.8	6.0	0.9
o6-78	*9*	*-177.5*	*4.8*	*5.9*	*0.6*
o6-79	31	178.6	4.3	7.2	0.3
o6-79	*23*	*-179.9*	*2.2*	*7.3*	*0.3*
o6-80	13	-179.9	3.9	7.0	1.3
o6-80	*12*	*179.8*	*3.8*	*7.2*	*1.1*
o6-81	13	179.5	3.2	6.5	0.5
o6-81	*14*	*-179.0*	*4.1*	*6.5*	*0.4*

Table A3.13. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*		*position 4*		*position 5*		*position 6*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-1														
Ala	9,210	0.94	1,289	0.79	568	0.35	1,742	1.06	2,080	1.27	1,491	0.91	2,040	1.25
Cys	1,325	0.93	270	1.13	314	1.32	90	0.38	88	0.37	277	1.16	286	1.20
Asp	9,628	1.42	602	0.53	3,520	3.11	1,094	0.97	2,023	1.79	1,991	1.76	398	0.35
Glu	10,380	1.37	709	0.56	673	0.53	1,844	1.46	3,719	2.94	2,894	2.29	541	0.43
Phe	3,856	0.83	1,015	1.31	163	0.21	604	0.78	448	0.58	678	0.88	948	1.23
Gly	5,282	0.62	1,222	0.85	911	0.64	526	0.37	825	0.58	800	0.56	998	0.70
His	2,395	0.88	349	0.77	574	1.27	270	0.60	418	0.92	483	1.07	301	0.67
Ile	5,355	0.82	1,665	1.52	142	0.13	753	0.69	465	0.43	604	0.55	1,726	1.58
Lys	5,612	0.89	773	0.74	480	0.46	1,079	1.03	1,296	1.23	928	0.88	1,056	1.01
Leu	9,818	0.93	2,639	1.50	376	0.21	1,664	0.94	842	0.48	1,263	0.72	3,034	1.72
Met	1,919	0.98	606	1.85	74	0.23	289	0.88	165	0.50	279	0.85	506	1.55
Asn	4,811	1.01	540	0.68	1,880	2.37	407	0.51	741	0.93	800	1.01	443	0.56
Pro	6,986	1.31	1,019	1.15	1,816	2.05	3,455	3.90	402	0.45	67	0.08	227	0.26
Gln	4,532	1.08	504	0.72	309	0.44	619	0.89	935	1.34	1,282	1.83	883	1.26
Arg	5,149	0.88	671	0.69	483	0.50	981	1.01	915	0.94	687	0.71	1,412	1.45
Ser	8,376	1.26	813	0.73	3,545	3.19	963	0.87	1,384	1.25	1,018	0.92	653	0.59
Thr	7,535	1.22	1,071	1.04	2,702	2.63	718	0.70	875	0.85	1,503	1.46	666	0.65
Val	6,787	0.83	1,916	1.41	212	0.16	1,086	0.80	665	0.49	1,131	0.83	1,777	1.31
Trp	1,521	0.94	360	1.34	67	0.25	300	1.11	235	0.87	201	0.75	358	1.33
Tyr	3,231	0.81	829	1.24	209	0.31	472	0.71	477	0.71	581	0.87	663	0.99
Other	556	0.99	182	1.93	26	0.28	88	0.93	46	0.49	86	0.91	128	1.36
type: o6-2														
Ala	3,549	0.96	728	1.19	592	0.96	254	0.41	588	0.96	690	1.12	697	1.14
Cys	558	1.04	159	1.78	62	0.69	148	1.65	18	0.20	44	0.49	127	1.42
Asp	4,273	1.68	330	0.78	583	1.37	1,488	3.50	447	1.05	788	1.86	637	1.50
Glu	3,543	1.25	260	0.55	543	1.14	260	0.55	668	1.41	1,135	2.39	677	1.43
Phe	1,483	0.85	387	1.33	235	0.81	106	0.37	171	0.59	212	0.73	372	1.28
Gly	2,405	0.75	584	1.09	318	0.59	427	0.79	213	0.40	277	0.52	586	1.09
His	1,119	1.10	126	0.74	191	1.13	323	1.90	88	0.52	185	1.09	206	1.21
Ile	1,420	0.58	444	1.08	196	0.48	79	0.19	233	0.57	191	0.47	277	0.68
Lys	2,141	0.91	247	0.63	535	1.36	189	0.48	456	1.16	434	1.10	280	0.71
Leu	2,581	0.65	712	1.08	425	0.64	181	0.27	437	0.66	356	0.54	470	0.71
Met	515	0.70	121	0.99	92	0.75	47	0.38	68	0.55	76	0.62	111	0.90
Asn	2,318	1.30	191	0.64	456	1.53	825	2.77	164	0.55	407	1.37	275	0.92
Pro	2,882	1.45	552	1.66	305	0.92	7	0.02	1,890	5.68	100	0.30	28	0.08
Gln	1,493	0.95	163	0.62	351	1.34	163	0.62	193	0.74	297	1.13	326	1.24
Arg	1,865	0.85	257	0.70	448	1.23	231	0.63	334	0.91	373	1.02	222	0.61
Ser	4,006	1.61	550	1.32	643	1.54	1,422	3.42	378	0.91	556	1.34	457	1.10
Thr	2,729	1.18	386	1.00	632	1.64	730	1.89	215	0.56	402	1.04	364	0.94
Val	1,964	0.64	510	1.00	246	0.48	97	0.19	316	0.62	257	0.50	538	1.05
Trp	516	0.85	100	0.99	88	0.87	37	0.37	88	0.87	87	0.86	116	1.15
Tyr	1,335	0.89	292	1.17	177	0.71	120	0.48	150	0.60	253	1.01	343	1.37
Other	163	0.77	44	1.25	25	0.71	9	0.25	28	0.79	23	0.65	34	0.96

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-3														
Ala	3,566	0.98	643	1.06	1,076	1.77	861	1.42	62	0.10	588	0.97	336	0.55
Cys	326	0.61	54	0.61	36	0.41	84	0.95	13	0.15	84	0.95	55	0.62
Asp	3,273	1.30	809	1.92	547	1.30	897	2.13	205	0.49	245	0.58	570	1.36
Glu	2,912	1.03	539	1.15	979	2.08	457	0.97	96	0.20	337	0.72	504	1.07
Phe	1,059	0.62	151	0.53	100	0.35	215	0.75	28	0.10	391	1.36	174	0.61
Gly	6,635	2.08	260	0.49	220	0.41	148	0.28	5,420	10.19	358	0.67	229	0.43
His	919	0.91	184	1.10	99	0.59	242	1.44	98	0.58	163	0.97	133	0.79
Ile	1,658	0.68	247	0.61	143	0.35	157	0.39	0	0.00	723	1.78	388	0.96
Lys	3,124	1.34	647	1.66	758	1.94	451	1.16	225	0.58	481	1.23	562	1.44
Leu	3,058	0.78	610	0.93	379	0.58	840	1.28	49	0.07	730	1.12	450	0.69
Met	550	0.75	94	0.77	79	0.65	162	1.33	25	0.21	120	0.99	70	0.58
Asn	2,338	1.32	415	1.41	281	0.95	599	2.03	458	1.55	295	1.00	290	0.98
Pro	1,208	0.61	105	0.32	339	1.03	6	0.02	0	0.00	1	0.00	757	2.30
Gln	1,716	1.10	323	1.24	400	1.54	397	1.53	119	0.46	255	0.98	222	0.86
Arg	2,487	1.15	552	1.53	520	1.44	437	1.21	141	0.39	414	1.15	423	1.17
Ser	2,252	0.91	395	0.96	599	1.45	381	0.92	88	0.21	286	0.69	503	1.22
Thr	1,813	0.79	492	1.29	216	0.57	227	0.60	8	0.02	311	0.82	559	1.47
Val	1,966	0.65	325	0.64	146	0.29	161	0.32	2	0.00	805	1.59	527	1.04
Trp	358	0.60	54	0.54	35	0.35	76	0.76	7	0.07	94	0.94	92	0.92
Tyr	1,032	0.69	146	0.59	92	0.37	231	0.93	23	0.09	341	1.38	199	0.80
Other	170	0.81	25	0.72	26	0.74	41	1.17	3	0.09	48	1.37	27	0.77
type: o6-4														
Ala	2,923	0.87	637	1.13	596	1.06	575	1.02	463	0.82	272	0.48	380	0.68
Cys	561	1.14	97	1.18	89	1.09	69	0.84	82	1.00	119	1.45	105	1.28
Asp	1,768	0.76	330	0.85	210	0.54	374	0.96	271	0.70	337	0.87	246	0.63
Glu	2,904	1.11	409	0.94	496	1.14	661	1.52	684	1.57	422	0.97	232	0.53
Phe	2,289	1.44	407	1.53	399	1.50	281	1.06	283	1.07	678	2.55	241	0.91
Gly	3,308	1.12	243	0.49	207	0.42	198	0.40	159	0.32	428	0.87	2,073	4.22
His	1,269	1.36	178	1.15	168	1.08	177	1.14	191	1.23	347	2.23	208	1.34
Ile	2,116	0.94	489	1.30	471	1.26	388	1.03	426	1.14	196	0.52	146	0.39
Lys	2,360	1.09	243	0.67	407	1.13	562	1.56	522	1.45	298	0.83	328	0.91
Leu	4,179	1.15	910	1.50	759	1.25	635	1.05	725	1.20	784	1.30	366	0.60
Met	789	1.17	182	1.62	172	1.53	132	1.17	123	1.09	113	1.01	67	0.60
Asn	1,650	1.01	226	0.83	239	0.88	248	0.91	238	0.87	380	1.39	319	1.17
Pro	637	0.35	123	0.40	108	0.35	46	0.15	18	0.06	10	0.03	332	1.09
Gln	1,473	1.02	207	0.86	229	0.95	331	1.38	291	1.21	208	0.87	207	0.86
Arg	2,086	1.04	269	0.80	379	1.13	494	1.48	451	1.35	233	0.70	260	0.78
Ser	1,722	0.75	284	0.75	334	0.88	268	0.70	254	0.67	247	0.65	335	0.88
Thr	1,933	0.91	274	0.78	276	0.78	333	0.94	413	1.17	444	1.26	193	0.55
Val	2,351	0.84	463	0.99	486	1.04	395	0.85	535	1.14	256	0.55	216	0.46
Trp	655	1.18	169	1.83	125	1.35	108	1.17	86	0.93	114	1.23	53	0.57
Tyr	2,075	1.51	365	1.59	351	1.53	232	1.01	289	1.26	627	2.74	211	0.92
Other	186	0.96	34	1.05	38	1.18	32	0.99	35	1.08	26	0.80	21	0.65

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-5														
Ala	1,075	0.76	197	0.84	250	1.06	163	0.69	182	0.77	129	0.55	154	0.65
Cys	189	0.92	36	1.05	18	0.52	36	1.05	50	1.46	13	0.38	36	1.05
Asp	1,077	1.10	189	1.16	204	1.25	238	1.46	72	0.44	270	1.66	104	0.64
Glu	1,054	0.97	154	0.85	296	1.62	197	1.08	129	0.71	171	0.94	107	0.59
Phe	684	1.03	111	1.00	60	0.54	133	1.19	207	1.86	55	0.49	118	1.06
Gly	1,091	0.88	226	1.10	103	0.50	64	0.31	400	1.94	62	0.30	236	1.14
His	493	1.26	80	1.23	59	0.91	108	1.66	128	1.96	43	0.66	75	1.15
Ile	696	0.74	87	0.55	71	0.45	122	0.78	142	0.90	78	0.50	196	1.25
Lys	1,034	1.14	126	0.83	305	2.02	213	1.41	126	0.83	162	1.07	102	0.67
Leu	1,286	0.85	302	1.19	156	0.61	234	0.92	213	0.84	103	0.41	278	1.10
Met	196	0.69	31	0.66	34	0.72	35	0.74	49	1.04	17	0.36	30	0.64
Asn	770	1.12	126	1.10	120	1.05	162	1.42	118	1.03	142	1.24	102	0.89
Pro	1,449	1.89	142	1.11	199	1.56	9	0.07	0	0.00	795	6.23	304	2.38
Gln	570	0.94	83	0.82	128	1.27	125	1.24	92	0.91	72	0.72	70	0.70
Arg	854	1.02	137	0.98	189	1.35	179	1.28	116	0.83	117	0.83	116	0.83
Ser	1,123	1.17	268	1.68	238	1.49	171	1.07	122	0.76	191	1.20	133	0.83
Thr	954	1.08	231	1.56	148	1.00	185	1.25	117	0.79	135	0.91	138	0.93
Val	995	0.85	98	0.50	85	0.43	158	0.81	228	1.16	121	0.62	305	1.56
Trp	198	0.85	31	0.80	25	0.64	38	0.98	51	1.31	17	0.44	36	0.93
Tyr	586	1.02	78	0.81	42	0.44	159	1.65	182	1.89	41	0.43	84	0.87
Other	78	0.96	9	0.66	12	0.89	13	0.96	18	1.33	8	0.59	18	1.33
type: o6-6														
Ala	1,359	0.88	305	1.19	282	1.10	322	1.25	163	0.64	80	0.31	207	0.81
Cys	188	0.84	43	1.15	20	0.53	22	0.59	52	1.39	11	0.29	40	1.07
Asp	761	0.71	98	0.55	155	0.87	132	0.74	135	0.76	119	0.67	122	0.69
Glu	1,355	1.14	213	1.07	386	1.94	403	2.03	136	0.69	68	0.34	149	0.75
Phe	704	0.97	125	1.03	73	0.60	95	0.78	223	1.84	21	0.17	167	1.38
Gly	2,323	1.72	100	0.45	89	0.40	78	0.35	80	0.36	1,793	7.98	183	0.81
His	377	0.89	67	0.94	62	0.87	68	0.96	70	0.99	57	0.80	53	0.75
Ile	854	0.83	241	1.41	157	0.92	137	0.80	82	0.48	2	0.01	235	1.37
Lys	1,375	1.39	197	1.20	326	1.98	295	1.79	121	0.73	230	1.40	206	1.25
Leu	1,724	1.04	423	1.53	284	1.03	303	1.10	374	1.35	35	0.13	305	1.10
Met	265	0.86	72	1.40	55	1.07	35	0.68	45	0.88	11	0.21	47	0.92
Asn	722	0.97	89	0.71	99	0.80	103	0.83	105	0.84	231	1.86	95	0.76
Pro	308	0.37	38	0.27	51	0.37	26	0.19	2	0.01	0	0.00	191	1.37
Gln	720	1.09	128	1.17	151	1.38	169	1.54	76	0.69	102	0.93	94	0.86
Arg	999	1.09	164	1.07	236	1.54	215	1.41	85	0.56	140	0.92	159	1.04
Ser	983	0.94	142	0.82	115	0.66	153	0.88	369	2.12	50	0.29	154	0.88
Thr	1,036	1.07	96	0.60	136	0.84	115	0.71	544	3.38	10	0.06	135	0.84
Val	959	0.75	245	1.15	169	0.79	186	0.87	75	0.35	5	0.02	279	1.31
Trp	207	0.82	47	1.11	31	0.73	27	0.64	47	1.11	3	0.07	52	1.23
Tyr	631	1.01	140	1.34	96	0.92	88	0.84	195	1.86	14	0.13	98	0.94
Other	78	0.88	15	1.02	15	1.02	16	1.08	9	0.61	6	0.41	17	1.15

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-7														
Ala	1,070	0.91	152	0.77	279	1.42	246	1.25	38	0.19	167	0.85	188	0.96
Cys	144	0.84	43	1.50	12	0.42	21	0.73	15	0.52	19	0.66	34	1.19
Asp	852	1.04	249	1.83	144	1.06	196	1.44	74	0.54	109	0.80	80	0.59
Glu	953	1.04	153	1.01	285	1.87	179	1.18	71	0.47	172	1.13	93	0.61
Phe	458	0.82	80	0.86	44	0.47	72	0.77	17	0.18	91	0.98	154	1.66
Gly	2,025	1.96	96	0.56	75	0.44	68	0.39	1,503	8.72	61	0.35	222	1.29
His	269	0.82	61	1.12	35	0.64	47	0.86	36	0.66	43	0.79	47	0.86
Ile	721	0.91	204	1.55	76	0.58	39	0.30	2	0.02	136	1.03	264	2.01
Lys	907	1.20	165	1.31	224	1.77	141	1.12	125	0.99	201	1.59	51	0.40
Leu	1,459	1.15	258	1.22	191	0.90	254	1.20	20	0.09	372	1.76	364	1.72
Met	234	0.99	48	1.22	46	1.17	50	1.27	7	0.18	49	1.24	34	0.86
Asn	666	1.16	66	0.69	89	0.93	183	1.92	187	1.96	85	0.89	56	0.59
Pro	164	0.26	20	0.19	99	0.93	2	0.02	0	0.00	0	0.00	43	0.40
Gln	531	1.05	79	0.94	141	1.68	109	1.30	46	0.55	114	1.36	42	0.50
Arg	758	1.08	123	1.05	200	1.71	148	1.26	95	0.81	137	1.17	55	0.47
Ser	690	0.86	100	0.75	118	0.88	223	1.67	34	0.25	131	0.98	84	0.63
Thr	564	0.76	71	0.57	92	0.74	175	1.42	3	0.02	127	1.03	96	0.78
Val	689	0.70	212	1.29	75	0.46	42	0.26	0	0.00	152	0.93	208	1.27
Trp	146	0.75	32	0.99	14	0.43	16	0.49	5	0.15	38	1.17	41	1.26
Tyr	392	0.81	65	0.81	43	0.54	72	0.90	9	0.11	76	0.95	127	1.58
Other	54	0.80	14	1.24	9	0.79	8	0.71	4	0.35	11	0.97	8	0.71
type: o6-8														
Ala	636	0.58	105	0.57	62	0.34	255	1.39	95	0.52	24	0.13	95	0.52
Cys	103	0.64	35	1.31	26	0.97	11	0.41	17	0.64	6	0.22	8	0.30
Asp	1,811	2.38	15	0.12	641	5.05	195	1.54	810	6.38	52	0.41	98	0.77
Glu	777	0.91	89	0.63	34	0.24	282	1.99	161	1.13	24	0.17	187	1.32
Phe	326	0.63	137	1.58	48	0.55	31	0.36	31	0.36	24	0.28	55	0.63
Gly	2,008	2.08	48	0.30	136	0.85	70	0.44	49	0.31	1,628	10.13	77	0.48
His	219	0.72	34	0.67	22	0.43	35	0.69	44	0.87	21	0.41	63	1.24
Ile	306	0.42	195	1.59	5	0.04	47	0.38	7	0.06	0	0.00	52	0.42
Lys	744	1.05	163	1.38	41	0.35	177	1.50	93	0.79	48	0.41	222	1.88
Leu	633	0.53	268	1.36	116	0.59	74	0.37	45	0.23	41	0.21	89	0.45
Met	122	0.55	44	1.20	22	0.60	10	0.27	14	0.38	14	0.38	18	0.49
Asn	1,028	1.93	33	0.37	299	3.36	88	0.99	306	3.44	110	1.24	192	2.16
Pro	358	0.60	57	0.57	19	0.19	278	2.80	1	0.01	0	0.00	3	0.03
Gln	415	0.88	70	0.89	31	0.40	67	0.85	92	1.17	26	0.33	129	1.64
Arg	629	0.96	147	1.35	62	0.57	109	1.00	57	0.52	48	0.44	206	1.89
Ser	931	1.25	54	0.43	236	1.90	231	1.86	162	1.30	43	0.35	205	1.65
Thr	867	1.25	210	1.82	278	2.41	70	0.61	81	0.70	1	0.01	227	1.97
Val	472	0.52	293	1.92	3	0.02	48	0.31	5	0.03	2	0.01	121	0.79
Trp	116	0.64	39	1.29	16	0.53	12	0.40	24	0.79	5	0.17	20	0.66
Tyr	274	0.61	82	1.10	33	0.44	37	0.49	42	0.56	17	0.23	63	0.84
Other	41	0.65	18	1.70	6	0.57	9	0.85	0	0.00	2	0.19	6	0.57

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-9														
Ala	439	0.45	73	0.45	100	0.62	85	0.53	15	0.09	90	0.56	76	0.47
Cys	118	0.84	33	1.40	24	1.02	13	0.55	2	0.09	11	0.47	35	1.49
Asp	705	1.05	51	0.46	57	0.51	419	3.76	73	0.65	38	0.34	67	0.60
Glu	743	0.99	169	1.36	140	1.12	122	0.98	27	0.22	184	1.48	101	0.81
Phe	400	0.88	127	1.67	106	1.39	42	0.55	16	0.21	41	0.54	68	0.89
Gly	1,921	2.27	96	0.68	34	0.24	247	1.75	1,456	10.31	41	0.29	47	0.33
His	315	1.18	59	1.32	70	1.57	69	1.55	20	0.45	51	1.14	46	1.03
Ile	555	0.86	133	1.24	197	1.83	0	0.00	1	0.01	102	0.95	122	1.13
Lys	857	1.38	87	0.84	152	1.47	100	0.97	25	0.24	367	3.55	126	1.22
Leu	660	0.63	178	1.03	131	0.75	30	0.17	7	0.04	101	0.58	213	1.23
Met	127	0.66	36	1.12	17	0.53	14	0.43	6	0.19	31	0.96	23	0.71
Asn	831	1.77	74	0.95	27	0.35	487	6.23	121	1.55	60	0.77	62	0.79
Pro	319	0.61	43	0.49	66	0.76	0	0.00	0	0.00	0	0.00	210	2.40
Gln	415	1.00	63	0.91	64	0.93	70	1.02	20	0.29	144	2.09	54	0.78
Arg	554	0.96	67	0.70	119	1.24	62	0.65	27	0.28	181	1.89	98	1.02
Ser	435	0.66	90	0.82	82	0.75	60	0.55	38	0.35	77	0.70	88	0.80
Thr	431	0.71	164	1.62	50	0.49	4	0.04	4	0.04	83	0.82	126	1.24
Val	824	1.02	176	1.31	277	2.06	5	0.04	1	0.01	186	1.39	179	1.33
Trp	145	0.91	33	1.24	33	1.24	4	0.15	2	0.08	15	0.56	58	2.18
Tyr	433	1.10	114	1.73	128	1.95	41	0.62	14	0.21	62	0.94	74	1.12
Other	35	0.63	11	1.19	3	0.32	3	0.32	2	0.22	12	1.29	4	0.43
type: o6-10														
Ala	728	0.90	141	1.05	168	1.25	157	1.17	107	0.80	96	0.71	59	0.44
Cys	119	1.01	33	1.68	16	0.82	5	0.25	20	1.02	43	2.19	2	0.10
Asp	547	0.98	108	1.16	62	0.67	96	1.03	51	0.55	196	2.11	34	0.37
Glu	535	0.86	38	0.37	105	1.01	180	1.73	118	1.13	62	0.60	32	0.31
Phe	441	1.16	81	1.27	58	0.91	32	0.50	85	1.34	173	2.72	12	0.19
Gly	247	0.35	46	0.39	52	0.44	39	0.33	21	0.18	37	0.31	52	0.44
His	311	1.39	43	1.16	40	1.08	39	1.05	64	1.72	111	2.98	14	0.38
Ile	425	0.79	143	1.59	67	0.75	45	0.50	113	1.26	42	0.47	15	0.17
Lys	609	1.18	47	0.54	146	1.69	182	2.11	127	1.47	53	0.61	54	0.63
Leu	898	1.03	300	2.07	173	1.19	127	0.88	176	1.21	98	0.68	24	0.17
Met	166	1.03	55	2.04	31	1.15	19	0.71	32	1.19	23	0.85	6	0.22
Asn	422	1.08	55	0.84	52	0.80	69	1.06	66	1.01	143	2.19	37	0.57
Pro	1,128	2.58	35	0.48	74	1.01	15	0.21	0	0.00	0	0.00	1,004	13.77
Gln	377	1.09	38	0.66	80	1.39	108	1.88	74	1.29	51	0.89	26	0.45
Arg	551	1.15	49	0.61	132	1.65	142	1.77	115	1.44	73	0.91	40	0.50
Ser	468	0.85	75	0.82	91	1.00	100	1.10	48	0.53	84	0.92	70	0.77
Thr	410	0.81	58	0.69	69	0.82	83	0.98	92	1.09	52	0.62	56	0.66
Val	458	0.68	115	1.03	66	0.59	60	0.54	137	1.22	66	0.59	14	0.13
Trp	141	1.06	44	1.99	21	0.95	14	0.63	22	0.99	31	1.40	9	0.41
Tyr	381	1.16	56	1.02	58	1.06	40	0.73	94	1.71	127	2.31	6	0.11
Other	34	0.73	6	0.78	5	0.65	14	1.81	4	0.52	5	0.65	0	0.00

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-11														
Ala	678	0.92	157	1.27	171	1.38	106	0.86	87	0.70	66	0.53	91	0.74
Cys	129	1.19	40	2.22	9	0.50	21	1.17	47	2.61	4	0.22	8	0.44
Asp	719	1.40	51	0.60	79	0.92	66	0.77	230	2.69	100	1.17	193	2.26
Glu	643	1.12	48	0.50	142	1.49	121	1.27	46	0.48	94	0.98	192	2.01
Phe	314	0.90	57	0.98	26	0.45	59	1.01	127	2.18	8	0.14	37	0.63
Gly	244	0.38	57	0.53	32	0.30	26	0.24	23	0.21	30	0.28	76	0.70
His	268	1.31	48	1.41	23	0.67	48	1.41	91	2.66	15	0.44	43	1.26
Ile	278	0.56	66	0.80	43	0.52	83	1.01	41	0.50	16	0.19	29	0.35
Lys	609	1.28	120	1.51	166	2.09	111	1.40	40	0.50	70	0.88	102	1.29
Leu	577	0.72	159	1.20	123	0.92	149	1.12	63	0.47	28	0.21	55	0.41
Met	114	0.77	32	1.29	17	0.69	31	1.25	14	0.57	4	0.16	16	0.65
Asn	511	1.42	51	0.85	53	0.88	73	1.22	177	2.95	29	0.48	128	2.14
Pro	892	2.22	52	0.78	66	0.99	3	0.04	0	0.00	765	11.43	6	0.09
Gln	375	1.18	62	1.17	96	1.82	78	1.48	42	0.80	34	0.64	63	1.19
Arg	526	1.19	98	1.33	133	1.81	129	1.75	30	0.41	56	0.76	80	1.09
Ser	614	1.22	115	1.37	96	1.15	57	0.68	161	1.92	58	0.69	127	1.52
Thr	378	0.81	65	0.84	62	0.80	85	1.10	52	0.67	32	0.41	82	1.06
Val	317	0.51	69	0.67	49	0.48	102	0.99	43	0.42	18	0.18	36	0.35
Trp	95	0.78	24	1.18	11	0.54	21	1.03	15	0.74	6	0.29	18	0.88
Tyr	317	1.05	58	1.15	34	0.67	63	1.25	105	2.08	5	0.10	52	1.03
Other	30	0.70	9	1.27	7	0.98	6	0.84	4	0.56	0	0.00	4	0.56
type: o6-12														
Ala	496	0.86	65	0.68	98	1.02	14	0.15	107	1.12	94	0.98	118	1.23
Cys	59	0.71	11	0.79	19	1.36	8	0.57	7	0.50	6	0.43	8	0.57
Asp	560	1.41	82	1.24	112	1.69	154	2.33	75	1.13	117	1.77	20	0.30
Glu	524	1.18	88	1.19	50	0.68	24	0.32	135	1.82	184	2.49	43	0.58
Phe	246	0.91	52	1.15	55	1.22	13	0.29	30	0.66	35	0.77	61	1.35
Gly	411	0.82	101	1.21	27	0.32	118	1.41	32	0.38	36	0.43	97	1.16
His	223	1.40	34	1.28	38	1.44	65	2.46	20	0.76	27	1.02	39	1.47
Ile	262	0.68	65	1.02	76	1.19	10	0.16	39	0.61	27	0.42	45	0.70
Lys	297	0.81	56	0.91	37	0.60	24	0.39	73	1.19	69	1.12	38	0.62
Leu	424	0.69	86	0.83	119	1.15	17	0.16	58	0.56	66	0.64	78	0.76
Met	94	0.82	18	0.94	28	1.46	3	0.16	12	0.63	13	0.68	20	1.04
Asn	450	1.62	32	0.69	79	1.70	146	3.15	26	0.56	74	1.59	93	2.00
Pro	432	1.39	102	1.97	66	1.27	1	0.02	243	4.68	15	0.29	5	0.10
Gln	195	0.79	35	0.86	42	1.03	14	0.34	18	0.44	58	1.42	28	0.68
Arg	270	0.79	46	0.81	46	0.81	16	0.28	47	0.83	69	1.21	46	0.81
Ser	552	1.42	47	0.72	47	0.72	226	3.48	62	0.96	77	1.19	93	1.43
Thr	525	1.46	47	0.78	29	0.48	225	3.74	45	0.75	68	1.13	111	1.85
Val	350	0.73	78	0.98	78	0.98	20	0.25	41	0.51	28	0.35	105	1.32
Trp	80	0.85	18	1.14	18	1.14	4	0.25	20	1.27	9	0.57	11	0.70
Tyr	214	0.91	47	1.20	47	1.20	12	0.31	17	0.44	39	1.00	52	1.33
Other	20	0.61	4	0.73	3	0.54	0	0.00	7	1.27	3	0.54	3	0.54

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-13														
Ala	332	0.47	62	0.53	66	0.56	7	0.06	65	0.56	61	0.52	71	0.61
Cys	81	0.79	21	1.23	8	0.47	0	0.00	6	0.35	23	1.35	23	1.35
Asp	507	1.05	27	0.33	329	4.07	57	0.70	15	0.19	56	0.69	23	0.28
Glu	472	0.87	92	1.02	90	1.00	21	0.23	158	1.75	80	0.88	31	0.34
Phe	262	0.79	88	1.59	28	0.51	10	0.18	27	0.49	43	0.78	66	1.19
Gly	1,264	2.06	12	0.12	125	1.22	1,074	10.49	14	0.14	11	0.11	28	0.27
His	225	1.16	50	1.55	56	1.73	17	0.53	39	1.21	23	0.71	40	1.24
Ile	492	1.05	159	2.04	0	0.00	0	0.00	66	0.85	89	1.14	178	2.28
Lys	643	1.43	108	1.44	70	0.93	14	0.19	313	4.17	81	1.08	57	0.76
Leu	515	0.68	117	0.93	18	0.14	3	0.02	62	0.49	171	1.36	144	1.14
Met	87	0.62	11	0.47	8	0.34	2	0.09	17	0.73	17	0.73	32	1.37
Asn	596	1.75	25	0.44	384	6.77	86	1.52	39	0.69	38	0.67	24	0.42
Pro	240	0.63	4	0.06	0	0.00	0	0.00	0	0.00	203	3.20	33	0.52
Gln	300	1.00	42	0.84	48	0.96	11	0.22	117	2.34	45	0.90	37	0.74
Arg	425	1.02	95	1.37	49	0.70	19	0.27	141	2.03	73	1.05	48	0.69
Ser	257	0.54	55	0.69	42	0.53	26	0.33	44	0.55	54	0.68	36	0.45
Thr	242	0.55	33	0.45	5	0.07	1	0.01	59	0.80	78	1.06	66	0.90
Val	745	1.28	237	2.44	4	0.04	1	0.01	124	1.27	126	1.30	253	2.60
Trp	104	0.90	20	1.04	1	0.05	1	0.05	11	0.57	47	2.44	24	1.25
Tyr	349	1.22	99	2.08	29	0.61	9	0.19	33	0.69	40	0.84	139	2.91
Other	28	0.69	4	0.59	1	0.15	2	0.30	11	1.63	2	0.30	8	1.19
type: o6-14														
Ala	654	1.25	143	1.64	70	0.80	27	0.31	91	1.04	107	1.23	216	2.48
Cys	101	1.33	19	1.50	20	1.57	20	1.57	9	0.71	15	1.18	18	1.42
Asp	274	0.76	17	0.28	52	0.86	28	0.46	96	1.59	71	1.18	10	0.17
Glu	302	0.75	18	0.27	61	0.90	18	0.27	92	1.36	81	1.20	32	0.47
Phe	197	0.80	57	1.38	36	0.87	6	0.15	32	0.78	41	0.99	25	0.61
Gly	1,408	3.08	181	2.37	82	1.07	764	10.01	58	0.76	182	2.38	141	1.85
His	143	0.99	15	0.62	51	2.12	5	0.21	19	0.79	31	1.29	22	0.91
Ile	252	0.72	49	0.84	70	1.20	2	0.03	73	1.25	21	0.36	37	0.64
Lys	236	0.70	30	0.54	45	0.80	12	0.21	61	1.09	31	0.55	57	1.02
Leu	377	0.67	119	1.27	50	0.53	5	0.05	64	0.68	64	0.68	75	0.80
Met	79	0.76	17	0.97	15	0.86	1	0.06	18	1.03	20	1.15	8	0.46
Asn	202	0.80	33	0.78	31	0.73	13	0.31	35	0.83	54	1.28	36	0.85
Pro	191	0.67	34	0.72	26	0.55	0	0.00	80	1.69	24	0.51	27	0.57
Gln	174	0.78	25	0.67	39	1.05	11	0.30	29	0.78	28	0.75	42	1.13
Arg	246	0.79	31	0.60	57	1.10	9	0.17	46	0.89	27	0.52	76	1.46
Ser	404	1.14	59	1.00	84	1.42	61	1.03	83	1.40	64	1.08	53	0.90
Thr	308	0.94	62	1.13	94	1.72	12	0.22	46	0.84	53	0.97	41	0.75
Val	256	0.59	57	0.79	63	0.87	2	0.03	39	0.54	42	0.58	53	0.73
Trp	80	0.93	13	0.91	24	1.67	4	0.28	12	0.84	14	0.97	13	0.91
Tyr	169	0.79	32	0.90	44	1.24	6	0.17	25	0.70	39	1.10	23	0.65
Other	37	1.23	4	0.80	1	0.20	9	1.79	7	1.40	6	1.20	10	1.99

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-15														
Ala	396	0.71	87	0.94	30	0.32	99	1.07	49	0.53	68	0.73	63	0.68
Cys	79	0.97	31	2.29	7	0.52	3	0.22	4	0.30	16	1.18	18	1.33
Asp	554	1.44	34	0.53	147	2.29	127	1.98	167	2.60	44	0.69	35	0.55
Glu	474	1.10	86	1.20	25	0.35	152	2.12	95	1.33	78	1.09	38	0.53
Phe	231	0.88	49	1.12	21	0.48	18	0.41	29	0.66	53	1.21	61	1.39
Gly	407	0.84	55	0.68	157	1.93	38	0.47	23	0.28	89	1.10	45	0.55
His	207	1.35	21	0.82	26	1.01	22	0.86	48	1.87	76	2.96	14	0.55
Ile	269	0.72	85	1.37	6	0.10	12	0.19	25	0.40	38	0.61	103	1.66
Lys	330	0.92	67	1.13	10	0.17	94	1.58	63	1.06	55	0.92	41	0.69
Leu	329	0.55	100	1.00	21	0.21	33	0.33	28	0.28	49	0.49	98	0.98
Met	61	0.55	15	0.81	5	0.27	6	0.32	10	0.54	13	0.70	12	0.65
Asn	418	1.55	28	0.62	136	3.02	47	1.05	113	2.51	53	1.18	41	0.91
Pro	339	1.12	48	0.96	0	0.00	185	3.68	3	0.06	0	0.00	103	2.05
Gln	182	0.77	22	0.56	16	0.40	30	0.76	51	1.29	40	1.01	23	0.58
Arg	276	0.83	44	0.80	19	0.34	48	0.87	63	1.14	60	1.09	42	0.76
Ser	683	1.81	71	1.13	225	3.58	91	1.45	107	1.70	110	1.75	79	1.26
Thr	599	1.72	64	1.10	203	3.49	41	0.70	110	1.89	86	1.48	95	1.63
Val	327	0.71	113	1.47	9	0.12	10	0.13	23	0.30	67	0.87	105	1.36
Trp	89	0.97	19	1.24	2	0.13	11	0.72	14	0.92	22	1.44	21	1.38
Tyr	212	0.93	36	0.95	14	0.37	12	0.32	51	1.35	57	1.51	42	1.11
Other	12	0.37	4	0.75	0	0.00	0	0.00	3	0.56	5	0.94	0	0.00
type: o6-16														
Ala	404	0.71	57	0.60	114	1.19	107	1.12	10	0.10	60	0.63	56	0.59
Cys	83	1.00	48	3.45	6	0.43	4	0.29	3	0.22	13	0.93	9	0.65
Asp	292	0.74	38	0.58	31	0.47	53	0.80	46	0.70	59	0.89	65	0.98
Glu	469	1.06	24	0.33	85	1.15	114	1.54	10	0.14	154	2.09	82	1.11
Phe	181	0.67	52	1.15	14	0.31	33	0.73	11	0.24	35	0.78	36	0.80
Gly	1,199	2.39	59	0.71	77	0.92	12	0.14	838	10.03	138	1.65	75	0.90
His	149	0.94	29	1.10	18	0.68	12	0.45	14	0.53	44	1.67	32	1.21
Ile	245	0.64	97	1.52	33	0.52	31	0.49	0	0.00	28	0.44	56	0.88
Lys	395	1.08	38	0.62	110	1.80	94	1.53	18	0.29	69	1.13	66	1.08
Leu	399	0.65	183	1.78	72	0.70	36	0.35	3	0.03	39	0.38	66	0.64
Met	73	0.64	15	0.79	15	0.79	12	0.63	6	0.31	20	1.05	5	0.26
Asn	229	0.82	32	0.69	29	0.63	27	0.58	78	1.68	27	0.58	36	0.78
Pro	658	2.12	73	1.41	170	3.29	337	6.51	0	0.00	0	0.00	78	1.51
Gln	230	0.94	40	0.98	44	1.08	43	1.05	10	0.25	71	1.74	22	0.54
Arg	262	0.77	30	0.53	72	1.27	47	0.83	17	0.30	39	0.69	57	1.00
Ser	389	1.00	53	0.82	65	1.00	42	0.65	24	0.37	100	1.54	105	1.62
Thr	313	0.87	57	0.95	38	0.63	28	0.47	9	0.15	74	1.23	107	1.79
Val	394	0.83	127	1.60	73	0.92	52	0.65	0	0.00	62	0.78	80	1.01
Trp	81	0.86	22	1.40	11	0.70	7	0.45	4	0.25	18	1.14	19	1.21
Tyr	194	0.83	29	0.74	32	0.82	18	0.46	8	0.21	52	1.34	55	1.41
Other	27	0.82	8	1.46	2	0.36	2	0.36	2	0.36	9	1.64	4	0.73

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-17														
Ala	380	0.82	81	1.05	106	1.38	79	1.03	43	0.56	4	0.05	67	0.87
Cys	49	0.73	15	1.34	3	0.27	12	1.07	8	0.71	1	0.09	10	0.89
Asp	272	0.85	53	1.00	65	1.22	48	0.90	27	0.51	6	0.11	73	1.37
Glu	394	1.10	63	1.06	114	1.92	101	1.70	42	0.71	1	0.02	73	1.23
Phe	228	1.04	41	1.13	18	0.49	27	0.74	94	2.58	0	0.00	48	1.32
Gly	998	2.47	29	0.43	25	0.37	12	0.18	27	0.40	857	12.72	48	0.71
His	149	1.17	30	1.41	24	1.13	21	0.99	55	2.58	2	0.09	17	0.80
Ile	250	0.81	56	1.09	30	0.58	52	1.01	65	1.26	0	0.00	47	0.91
Lys	311	1.05	41	0.83	81	1.64	90	1.82	43	0.87	3	0.06	53	1.07
Leu	463	0.93	115	1.39	66	0.80	93	1.12	120	1.45	2	0.02	67	0.81
Met	69	0.75	17	1.10	10	0.65	14	0.91	16	1.04	1	0.06	11	0.71
Asn	176	0.79	33	0.88	36	0.96	25	0.67	39	1.04	3	0.08	40	1.07
Pro	158	0.63	23	0.55	35	0.84	3	0.07	2	0.05	0	0.00	95	2.28
Gln	182	0.92	35	1.06	53	1.61	38	1.16	31	0.94	1	0.03	24	0.73
Arg	298	1.08	61	1.33	79	1.72	77	1.68	41	0.90	5	0.11	35	0.76
Ser	197	0.63	51	0.98	43	0.82	32	0.61	24	0.46	6	0.11	41	0.79
Thr	217	0.75	47	0.97	37	0.77	54	1.12	40	0.83	3	0.06	36	0.74
Val	282	0.73	66	1.03	37	0.58	71	1.11	55	0.86	0	0.00	53	0.83
Trp	73	0.96	13	1.03	12	0.95	13	1.03	25	1.97	0	0.00	10	0.79
Tyr	213	1.13	22	0.70	18	0.57	31	0.99	95	3.03	1	0.03	46	1.46
Other	17	0.64	4	0.90	4	0.90	3	0.68	4	0.90	0	0.00	2	0.45
type: o6-18														
Ala	418	0.94	91	1.23	63	0.85	78	1.05	71	0.96	13	0.18	102	1.37
Cys	64	0.99	24	2.22	5	0.46	7	0.65	3	0.28	3	0.28	22	2.03
Asp	279	0.91	42	0.82	68	1.32	26	0.51	51	0.99	61	1.19	31	0.60
Glu	318	0.92	65	1.13	50	0.87	68	1.18	94	1.64	12	0.21	29	0.51
Phe	169	0.80	29	0.83	27	0.77	25	0.71	14	0.40	17	0.48	57	1.62
Gly	774	1.99	51	0.78	35	0.54	51	0.78	10	0.15	533	8.20	94	1.45
His	91	0.74	16	0.78	18	0.88	17	0.83	8	0.39	18	0.88	14	0.68
Ile	215	0.72	24	0.48	36	0.73	84	1.69	18	0.36	0	0.00	53	1.07
Lys	266	0.93	45	0.94	57	1.20	45	0.94	59	1.24	20	0.42	40	0.84
Leu	329	0.69	83	1.04	77	0.96	76	0.95	24	0.30	14	0.18	55	0.69
Met	73	0.82	20	1.35	13	0.88	13	0.88	7	0.47	5	0.34	15	1.01
Asn	262	1.21	44	1.22	46	1.28	23	0.64	29	0.81	94	2.61	26	0.72
Pro	435	1.80	56	1.39	49	1.22	4	0.10	311	7.73	0	0.00	15	0.37
Gln	201	1.06	29	0.91	37	1.17	54	1.70	27	0.85	13	0.41	41	1.29
Arg	192	0.72	26	0.59	50	1.13	52	1.18	33	0.75	11	0.25	20	0.45
Ser	361	1.20	93	1.85	82	1.63	45	0.89	39	0.77	25	0.50	77	1.53
Thr	281	1.01	51	1.09	83	1.78	74	1.59	22	0.47	3	0.06	48	1.03
Val	248	0.67	34	0.55	32	0.52	74	1.20	27	0.44	0	0.00	81	1.31
Trp	53	0.72	10	0.82	11	0.90	10	0.82	5	0.41	4	0.33	13	1.06
Tyr	142	0.78	27	0.89	24	0.79	33	1.09	11	0.36	17	0.56	30	0.99
Other	13	0.51	4	0.94	1	0.23	5	1.17	1	0.23	1	0.23	1	0.23

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-19														
Ala	365	0.87	55	0.79	51	0.73	110	1.57	63	0.90	42	0.60	44	0.63
Cys	54	0.88	8	0.79	7	0.69	23	2.26	2	0.20	7	0.69	7	0.69
Asp	378	1.30	62	1.28	65	1.34	40	0.83	115	2.38	66	1.37	30	0.62
Glu	321	0.99	44	0.81	46	0.85	37	0.69	89	1.65	67	1.24	38	0.70
Phe	170	0.86	35	1.06	27	0.82	18	0.55	17	0.51	56	1.70	17	0.51
Gly	312	0.85	68	1.11	33	0.54	120	1.96	28	0.46	28	0.46	35	0.57
His	144	1.24	27	1.40	30	1.55	10	0.52	15	0.78	31	1.61	31	1.61
Ile	150	0.54	24	0.51	24	0.51	9	0.19	10	0.21	42	0.90	41	0.88
Lys	307	1.14	38	0.85	65	1.45	29	0.65	71	1.58	48	1.07	56	1.25
Leu	294	0.65	67	0.89	61	0.81	42	0.56	34	0.45	48	0.64	42	0.56
Met	58	0.69	11	0.79	9	0.64	8	0.57	10	0.72	7	0.50	13	0.93
Asn	261	1.28	77	2.27	38	1.12	28	0.83	26	0.77	69	2.04	23	0.68
Pro	401	1.77	49	1.29	31	0.82	7	0.18	123	3.25	0	0.00	191	5.05
Gln	199	1.11	25	0.84	40	1.34	34	1.14	25	0.84	40	1.34	35	1.17
Arg	226	0.91	38	0.91	54	1.30	29	0.70	35	0.84	39	0.94	31	0.75
Ser	457	1.61	65	1.37	81	1.71	133	2.81	59	1.25	52	1.10	67	1.41
Thr	343	1.30	51	1.16	73	1.66	80	1.82	40	0.91	58	1.32	41	0.93
Val	222	0.64	37	0.64	42	0.72	20	0.34	33	0.57	51	0.88	39	0.67
Trp	54	0.78	12	1.04	9	0.78	8	0.70	5	0.43	9	0.78	11	0.96
Tyr	148	0.87	18	0.63	23	0.81	27	0.95	11	0.39	50	1.75	19	0.67
Other	14	0.58	2	0.50	4	1.00	1	0.25	2	0.50	3	0.75	2	0.50
type: o6-20														
Ala	640	1.28	58	0.70	16	0.19	103	1.24	104	1.25	203	2.44	156	1.87
Cys	65	0.89	14	1.15	6	0.49	7	0.58	15	1.24	8	0.66	15	1.24
Asp	252	0.73	38	0.66	19	0.33	51	0.89	89	1.55	10	0.17	45	0.78
Glu	401	1.04	42	0.65	15	0.23	99	1.54	124	1.93	69	1.07	52	0.81
Phe	179	0.76	36	0.91	1	0.03	35	0.89	28	0.71	32	0.81	47	1.19
Gly	1,356	3.11	173	2.37	810	11.12	32	0.44	118	1.62	177	2.43	46	0.63
His	106	0.77	46	2.00	1	0.04	15	0.65	14	0.61	19	0.83	11	0.48
Ile	269	0.81	53	0.95	0	0.00	82	1.48	22	0.40	24	0.43	88	1.58
Lys	257	0.80	45	0.84	5	0.09	54	1.01	49	0.92	57	1.07	47	0.88
Leu	383	0.71	80	0.89	0	0.00	84	0.94	55	0.61	66	0.74	98	1.09
Met	80	0.80	17	1.02	0	0.00	22	1.32	14	0.84	2	0.12	25	1.50
Asn	169	0.70	21	0.52	10	0.25	27	0.67	31	0.77	46	1.14	34	0.84
Pro	139	0.51	25	0.55	0	0.00	63	1.40	31	0.69	15	0.33	5	0.11
Gln	176	0.83	38	1.07	8	0.22	29	0.82	42	1.18	26	0.73	33	0.93
Arg	245	0.83	50	1.01	5	0.10	51	1.03	35	0.71	64	1.29	40	0.81
Ser	298	0.88	55	0.97	50	0.89	63	1.12	41	0.73	40	0.71	49	0.87
Thr	271	0.87	58	1.11	14	0.27	41	0.78	77	1.47	29	0.55	52	0.99
Val	246	0.59	52	0.75	0	0.00	54	0.78	40	0.58	39	0.56	61	0.88
Trp	81	0.99	16	1.17	1	0.07	23	1.68	10	0.73	15	1.09	16	1.17
Tyr	167	0.82	46	1.35	3	0.09	29	0.85	27	0.79	21	0.62	41	1.21
Other	34	1.18	6	1.25	5	1.04	5	1.04	3	0.63	7	1.46	8	1.67

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-21														
Ala	406	1.06	92	1.44	74	1.16	69	1.08	53	0.83	44	0.69	74	1.16
Cys	63	1.13	13	1.40	6	0.64	10	1.07	20	2.15	6	0.64	8	0.86
Asp	376	1.42	45	1.02	44	1.00	58	1.31	20	0.45	167	3.78	42	0.95
Glu	256	0.86	22	0.45	66	1.34	54	1.09	5	0.10	54	1.09	55	1.11
Phe	148	0.82	40	1.33	20	0.66	23	0.76	13	0.43	11	0.36	41	1.36
Gly	305	0.91	50	0.89	26	0.47	27	0.48	110	1.97	37	0.66	55	0.98
His	100	0.94	13	0.74	9	0.51	34	1.93	11	0.62	10	0.57	23	1.30
Ile	148	0.58	42	0.99	27	0.63	19	0.45	6	0.14	10	0.23	44	1.03
Lys	247	1.01	37	0.90	61	1.49	59	1.44	13	0.32	35	0.85	42	1.03
Leu	378	0.92	91	1.32	104	1.51	58	0.84	18	0.26	29	0.42	78	1.13
Met	56	0.73	16	1.25	15	1.17	9	0.70	5	0.39	2	0.16	9	0.70
Asn	237	1.28	32	1.03	38	1.23	63	2.03	20	0.65	48	1.55	36	1.16
Pro	174	0.84	31	0.90	50	1.45	2	0.06	1	0.03	85	2.46	5	0.14
Gln	146	0.89	25	0.92	27	0.99	43	1.58	13	0.48	16	0.59	22	0.81
Arg	198	0.87	32	0.84	44	1.16	41	1.08	15	0.39	27	0.71	39	1.03
Ser	484	1.86	47	1.09	48	1.11	54	1.25	210	4.85	83	1.92	42	0.97
Thr	333	1.39	22	0.55	31	0.77	43	1.07	161	4.02	42	1.05	34	0.85
Val	179	0.56	43	0.81	19	0.36	27	0.51	32	0.60	18	0.34	40	0.75
Trp	59	0.94	19	1.81	8	0.76	8	0.76	2	0.19	6	0.57	16	1.52
Tyr	147	0.94	29	1.11	20	0.77	37	1.42	15	0.58	11	0.42	35	1.34
Other	18	0.82	2	0.54	6	1.63	5	1.36	0	0.00	2	0.54	3	0.82
type: o6-22														
Ala	186	0.47	33	0.50	32	0.49	13	0.20	37	0.56	30	0.46	41	0.62
Cys	53	0.92	8	0.84	17	1.77	1	0.10	3	0.31	6	0.63	18	1.88
Asp	382	1.40	11	0.24	67	1.47	50	1.10	193	4.24	49	1.08	12	0.26
Glu	295	0.97	32	0.63	72	1.42	22	0.43	75	1.48	65	1.28	29	0.57
Phe	153	0.82	32	1.03	39	1.26	2	0.06	6	0.19	38	1.22	36	1.16
Gly	617	1.79	18	0.31	2	0.03	551	9.58	23	0.40	8	0.14	15	0.26
His	107	0.98	21	1.16	29	1.60	10	0.55	5	0.28	34	1.87	8	0.44
Ile	232	0.88	54	1.23	67	1.53	3	0.07	5	0.11	36	0.82	67	1.53
Lys	324	1.28	93	2.20	54	1.28	28	0.66	45	1.07	75	1.78	29	0.69
Leu	280	0.66	49	0.69	49	0.69	3	0.04	11	0.16	65	0.92	103	1.46
Met	57	0.72	14	1.06	9	0.68	3	0.23	4	0.30	14	1.06	13	0.99
Asn	253	1.32	8	0.25	36	1.13	32	1.00	137	4.30	24	0.75	16	0.50
Pro	79	0.37	26	0.73	0	0.00	0	0.00	34	0.95	0	0.00	19	0.53
Gln	178	1.06	47	1.67	33	1.17	10	0.36	23	0.82	41	1.46	24	0.85
Arg	281	1.20	84	2.15	43	1.10	19	0.49	30	0.77	59	1.51	46	1.18
Ser	244	0.91	35	0.79	48	1.08	10	0.22	75	1.68	40	0.90	36	0.81
Thr	259	1.05	57	1.38	23	0.56	0	0.00	45	1.09	73	1.77	61	1.48
Val	372	1.13	89	1.63	101	1.85	1	0.02	4	0.07	52	0.95	125	2.29
Trp	59	0.91	14	1.29	10	0.92	1	0.09	0	0.00	16	1.48	18	1.66
Tyr	159	0.99	34	1.27	30	1.12	5	0.19	10	0.37	35	1.31	45	1.68
Other	20	0.88	6	1.59	4	1.06	1	0.26	0	0.00	5	1.32	4	1.06

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-23														
Ala	269	0.92	57	1.16	19	0.39	15	0.31	43	0.88	64	1.30	71	1.45
Cys	31	0.72	3	0.42	4	0.56	3	0.42	4	0.56	3	0.42	14	1.96
Asp	291	1.43	52	1.53	25	0.74	84	2.47	34	1.00	45	1.33	51	1.50
Glu	264	1.16	41	1.08	13	0.34	19	0.50	50	1.32	81	2.14	60	1.58
Phe	119	0.86	33	1.42	12	0.52	3	0.13	20	0.86	25	1.08	26	1.12
Gly	517	2.01	29	0.68	349	8.13	53	1.23	10	0.23	36	0.84	40	0.93
His	87	1.07	16	1.18	15	1.11	17	1.25	12	0.88	10	0.74	17	1.25
Ile	77	0.39	21	0.64	3	0.09	3	0.09	22	0.67	15	0.46	13	0.40
Lys	164	0.87	36	1.14	21	0.67	11	0.35	40	1.27	31	0.98	25	0.79
Leu	164	0.52	22	0.42	14	0.27	14	0.27	49	0.93	43	0.81	22	0.42
Met	36	0.61	7	0.71	1	0.10	1	0.10	6	0.61	9	0.92	12	1.22
Asn	188	1.32	26	1.09	46	1.93	51	2.14	7	0.29	33	1.39	25	1.05
Pro	174	1.09	46	1.73	0	0.00	0	0.00	118	4.44	9	0.34	1	0.04
Gln	97	0.77	13	0.62	9	0.43	12	0.57	13	0.62	24	1.14	26	1.24
Arg	118	0.67	20	0.69	14	0.48	16	0.55	24	0.82	19	0.65	25	0.86
Ser	318	1.60	47	1.41	12	0.36	123	3.70	41	1.23	34	1.02	61	1.83
Thr	264	1.43	45	1.46	1	0.03	128	4.16	27	0.88	37	1.20	26	0.84
Val	89	0.36	20	0.49	0	0.00	5	0.12	28	0.69	22	0.54	14	0.34
Trp	45	0.93	9	1.11	7	0.87	0	0.00	12	1.49	7	0.87	10	1.24
Tyr	104	0.87	27	1.35	6	0.30	11	0.55	10	0.50	23	1.15	27	1.35
Other	10	0.59	1	0.35	0	0.00	2	0.71	1	0.35	1	0.35	5	1.77
type: o6-24														
Ala	166	0.67	29	0.71	41	1.00	8	0.19	29	0.71	23	0.56	36	0.88
Cys	33	0.92	13	2.17	3	0.50	1	0.17	7	1.17	0	0.00	9	1.50
Asp	162	0.95	11	0.39	34	1.20	39	1.37	10	0.35	52	1.83	16	0.56
Glu	136	0.71	19	0.60	36	1.13	7	0.22	29	0.91	34	1.07	11	0.35
Phe	90	0.77	44	2.27	9	0.46	8	0.41	11	0.57	5	0.26	13	0.67
Gly	401	1.86	36	1.00	9	0.25	278	7.73	49	1.36	11	0.31	18	0.50
His	67	0.98	13	1.14	10	0.88	16	1.41	14	1.23	10	0.88	4	0.35
Ile	131	0.80	23	0.84	24	0.88	1	0.04	23	0.84	23	0.84	37	1.35
Lys	229	1.45	24	0.91	54	2.05	8	0.30	102	3.87	33	1.25	8	0.30
Leu	202	0.76	71	1.61	31	0.70	6	0.14	11	0.25	28	0.63	55	1.24
Met	38	0.77	11	1.34	9	1.10	3	0.37	7	0.85	1	0.12	7	0.85
Asn	151	1.26	12	0.60	22	1.10	63	3.16	11	0.55	19	0.95	24	1.21
Pro	212	1.59	26	1.17	81	3.64	0	0.00	0	0.00	55	2.47	50	2.25
Gln	85	0.81	8	0.46	16	0.91	9	0.51	26	1.48	10	0.57	16	0.91
Arg	157	1.07	22	0.90	21	0.86	11	0.45	44	1.80	37	1.51	22	0.90
Ser	122	0.73	17	0.61	21	0.75	11	0.39	21	0.75	36	1.29	16	0.57
Thr	147	0.95	28	1.09	15	0.58	1	0.04	28	1.09	36	1.40	39	1.51
Val	229	1.12	42	1.23	25	0.73	0	0.00	30	0.88	48	1.41	84	2.46
Trp	24	0.59	5	0.74	3	0.44	2	0.30	2	0.30	5	0.74	7	1.03
Tyr	81	0.81	24	1.43	14	0.84	6	0.36	19	1.13	12	0.72	6	0.36
Other	5	0.35	0	0.00	0	0.00	0	0.00	5	2.12	0	0.00	0	0.00

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-25														
Ala	128	0.47	32	0.71	8	0.18	23	0.51	22	0.49	23	0.51	20	0.44
Cys	35	0.89	9	1.37	0	0.00	4	0.61	4	0.61	11	1.67	7	1.06
Asp	250	1.34	42	1.35	29	0.93	126	4.04	29	0.93	16	0.51	8	0.26
Glu	192	0.92	42	1.20	12	0.34	56	1.61	45	1.29	20	0.57	17	0.49
Phe	120	0.94	29	1.36	2	0.09	8	0.38	21	0.99	23	1.08	37	1.74
Gly	457	1.93	8	0.20	399	10.11	19	0.48	10	0.25	11	0.28	10	0.25
His	79	1.06	22	1.76	6	0.48	5	0.40	29	2.33	3	0.24	14	1.12
Ile	172	0.95	39	1.29	3	0.10	5	0.17	16	0.53	36	1.20	73	2.42
Lys	155	0.89	28	0.97	13	0.45	30	1.04	49	1.69	21	0.73	14	0.48
Leu	247	0.85	41	0.84	3	0.06	7	0.14	44	0.91	73	1.50	79	1.63
Met	42	0.78	11	1.22	2	0.22	6	0.66	4	0.44	9	1.00	10	1.11
Asn	172	1.31	20	0.91	23	1.05	83	3.79	20	0.91	16	0.73	10	0.46
Pro	89	0.61	4	0.16	0	0.00	26	1.06	0	0.00	44	1.80	15	0.61
Gln	117	1.01	25	1.30	6	0.31	19	0.99	40	2.08	19	0.99	8	0.42
Arg	143	0.89	22	0.82	4	0.15	19	0.71	43	1.60	30	1.12	25	0.93
Ser	157	0.86	29	0.95	8	0.26	47	1.54	34	1.11	18	0.59	21	0.69
Thr	162	0.95	17	0.60	0	0.00	31	1.09	54	1.91	38	1.34	22	0.78
Val	266	1.18	73	1.95	1	0.03	1	0.03	26	0.69	73	1.95	92	2.45
Trp	34	0.76	8	1.08	1	0.13	0	0.00	7	0.94	14	1.88	4	0.54
Tyr	117	1.06	20	1.09	3	0.16	8	0.43	26	1.41	24	1.30	36	1.96
Other	16	1.03	4	1.54	2	0.77	2	0.77	2	0.77	3	1.16	3	1.16
type: o6-26														
Ala	192	0.78	30	0.74	32	0.78	39	0.96	3	0.07	50	1.23	38	0.93
Cys	36	1.01	8	1.35	5	0.84	1	0.17	3	0.50	11	1.85	8	1.35
Asp	121	0.71	7	0.25	22	0.78	23	0.81	18	0.64	20	0.71	31	1.10
Glu	199	1.05	18	0.57	31	0.98	46	1.46	8	0.25	38	1.20	58	1.84
Phe	116	1.00	38	1.97	12	0.62	8	0.41	5	0.26	29	1.50	24	1.24
Gly	508	2.37	57	1.60	18	0.50	15	0.42	343	9.60	18	0.50	57	1.60
His	57	0.84	12	1.06	8	0.71	8	0.71	6	0.53	10	0.89	13	1.15
Ile	94	0.58	25	0.92	27	0.99	7	0.26	6	0.22	13	0.48	16	0.59
Lys	142	0.90	15	0.57	35	1.34	40	1.53	9	0.34	32	1.22	11	0.42
Leu	187	0.71	54	1.23	40	0.91	15	0.34	6	0.14	29	0.66	43	0.98
Met	41	0.84	13	1.59	4	0.49	3	0.37	2	0.24	12	1.47	7	0.86
Asn	87	0.73	15	0.76	13	0.66	8	0.40	17	0.86	21	1.06	13	0.66
Pro	301	2.27	34	1.54	78	3.53	169	7.64	0	0.00	0	0.00	20	0.90
Gln	108	1.03	16	0.92	20	1.15	14	0.80	6	0.34	34	1.95	18	1.03
Arg	110	0.76	19	0.78	31	1.28	18	0.74	11	0.45	16	0.66	15	0.62
Ser	144	0.87	24	0.87	19	0.69	25	0.90	7	0.25	42	1.52	27	0.98
Thr	116	0.76	22	0.86	24	0.94	3	0.12	3	0.12	42	1.64	22	0.86
Val	156	0.77	40	1.18	32	0.94	19	0.56	3	0.09	26	0.77	36	1.06
Trp	27	0.67	9	1.34	3	0.45	4	0.60	4	0.60	4	0.60	3	0.45
Tyr	97	0.97	17	1.02	18	1.08	9	0.54	15	0.90	25	1.50	13	0.78
Other	11	0.78	2	0.85	3	1.28	1	0.43	0	0.00	3	1.28	2	0.85

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-27														
Ala	138	0.62	26	0.70	16	0.43	43	1.16	33	0.89	1	0.03	19	0.51
Cys	18	0.55	3	0.55	5	0.92	2	0.37	3	0.55	0	0.00	5	0.92
Asp	158	1.02	12	0.47	36	1.40	29	1.13	55	2.14	2	0.08	24	0.93
Glu	170	0.99	43	1.49	3	0.10	49	1.70	48	1.67	3	0.10	24	0.83
Phe	70	0.66	18	1.02	11	0.63	10	0.57	14	0.80	2	0.11	15	0.85
Gly	567	2.90	21	0.64	67	2.06	10	0.31	21	0.64	407	12.50	41	1.26
His	48	0.78	8	0.78	14	1.36	2	0.19	13	1.26	0	0.00	11	1.07
Ile	74	0.50	21	0.85	3	0.12	11	0.44	7	0.28	1	0.04	31	1.25
Lys	115	0.80	30	1.26	2	0.08	23	0.96	38	1.59	2	0.08	20	0.84
Leu	129	0.54	31	0.77	18	0.45	15	0.37	25	0.62	1	0.02	39	0.97
Met	22	0.49	4	0.54	3	0.40	2	0.27	8	1.08	0	0.00	5	0.67
Asn	99	0.92	18	1.00	23	1.27	12	0.67	30	1.66	1	0.06	15	0.83
Pro	192	1.59	26	1.29	24	1.19	101	5.01	3	0.15	0	0.00	38	1.88
Gln	82	0.86	31	1.95	1	0.06	12	0.75	23	1.45	1	0.06	14	0.88
Arg	110	0.83	41	1.85	8	0.36	20	0.90	24	1.08	2	0.09	15	0.68
Ser	190	1.26	25	0.99	63	2.50	43	1.70	32	1.27	4	0.16	23	0.91
Thr	203	1.45	27	1.16	108	4.62	14	0.60	25	1.07	3	0.13	26	1.11
Val	105	0.57	28	0.90	5	0.16	25	0.81	5	0.16	1	0.03	41	1.32
Trp	29	0.79	5	0.82	5	0.82	6	0.98	1	0.16	0	0.00	12	1.96
Tyr	72	0.79	13	0.86	18	1.19	3	0.20	23	1.52	2	0.13	13	0.86
Other	7	0.55	2	0.93	0	0.00	1	0.47	2	0.93	0	0.00	2	0.93
type: o6-28														
Ala	123	0.77	15	0.56	28	1.04	32	1.19	13	0.49	17	0.63	18	0.67
Cys	29	1.24	15	3.84	1	0.26	2	0.51	3	0.77	5	1.28	3	0.77
Asp	171	1.54	6	0.32	25	1.35	26	1.40	63	3.40	30	1.62	21	1.13
Glu	149	1.20	3	0.14	16	0.77	28	1.35	26	1.25	49	2.36	27	1.30
Phe	45	0.59	14	1.11	3	0.24	9	0.71	7	0.55	3	0.24	9	0.71
Gly	174	1.24	29	1.24	46	1.96	10	0.43	24	1.02	55	2.34	10	0.43
His	49	1.10	10	1.35	10	1.35	5	0.67	18	2.43	2	0.27	4	0.54
Ile	57	0.53	22	1.23	0	0.00	7	0.39	6	0.34	6	0.34	16	0.89
Lys	97	0.94	10	0.58	17	0.99	25	1.45	11	0.64	15	0.87	19	1.10
Leu	127	0.73	62	2.15	14	0.49	10	0.35	7	0.24	13	0.45	21	0.73
Met	29	0.90	11	2.05	4	0.75	5	0.93	1	0.19	4	0.75	4	0.75
Asn	111	1.42	6	0.46	16	1.23	13	1.00	51	3.92	6	0.46	19	1.46
Pro	176	2.02	23	1.58	59	4.06	69	4.75	2	0.14	0	0.00	23	1.58
Gln	63	0.92	5	0.44	7	0.61	12	1.05	6	0.52	24	2.10	9	0.79
Arg	78	0.82	14	0.88	17	1.07	10	0.63	4	0.25	6	0.38	27	1.69
Ser	134	1.23	10	0.55	25	1.38	23	1.27	24	1.32	31	1.71	21	1.16
Thr	109	1.08	18	1.07	18	1.07	8	0.48	26	1.54	11	0.65	28	1.66
Val	84	0.63	22	0.99	2	0.09	9	0.40	3	0.13	27	1.21	21	0.94
Trp	21	0.79	3	0.68	0	0.00	7	1.59	5	1.13	4	0.91	2	0.45
Tyr	37	0.56	10	0.91	3	0.27	2	0.18	10	0.91	3	0.27	9	0.82
Other	9	0.97	4	2.59	1	0.65	0	0.00	2	1.30	1	0.65	1	0.65

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-29														
Ala	105	0.80	24	1.09	17	0.77	22	1.00	12	0.55	17	0.77	13	0.59
Cys	23	1.20	4	1.25	0	0.00	1	0.31	6	1.87	5	1.56	7	2.18
Asp	82	0.90	14	0.92	17	1.12	10	0.66	11	0.72	3	0.20	27	1.77
Glu	100	0.98	15	0.88	17	1.00	27	1.59	16	0.94	7	0.41	18	1.06
Phe	70	1.12	10	0.96	12	1.15	5	0.48	30	2.89	9	0.87	4	0.38
Gly	172	1.49	8	0.42	5	0.26	3	0.16	6	0.31	130	6.75	20	1.04
His	44	1.21	5	0.82	2	0.33	9	1.48	22	3.62	3	0.49	3	0.49
Ile	72	0.82	22	1.50	17	1.16	17	1.16	12	0.82	1	0.07	3	0.20
Lys	88	1.04	6	0.43	20	1.42	32	2.27	10	0.71	5	0.35	15	1.06
Leu	129	0.91	38	1.60	30	1.27	29	1.22	13	0.55	12	0.51	7	0.30
Met	21	0.80	4	0.91	7	1.59	3	0.68	2	0.45	3	0.68	2	0.45
Asn	75	1.17	11	1.03	14	1.31	8	0.75	18	1.69	2	0.19	22	2.06
Pro	36	0.50	8	0.67	7	0.59	2	0.17	0	0.00	0	0.00	19	1.59
Gln	54	0.96	9	0.96	17	1.81	14	1.49	5	0.53	4	0.43	5	0.53
Arg	85	1.08	12	0.92	26	1.99	25	1.91	8	0.61	2	0.15	12	0.92
Ser	87	0.97	8	0.54	16	1.07	11	0.74	11	0.74	17	1.14	24	1.61
Thr	98	1.18	6	0.43	11	0.80	19	1.38	13	0.94	23	1.67	26	1.88
Val	83	0.76	18	0.98	14	0.77	11	0.60	19	1.04	4	0.22	17	0.93
Trp	28	1.29	12	3.31	3	0.83	2	0.55	5	1.38	3	0.83	3	0.83
Tyr	76	1.41	19	2.12	4	0.45	4	0.45	36	4.01	6	0.67	7	0.78
Other	8	1.05	3	2.37	0	0.00	2	1.58	1	0.79	0	0.00	2	1.58
type: o6-30														
Ala	112	0.84	19	0.86	23	1.04	8	0.36	11	0.50	23	1.04	28	1.26
Cys	12	0.62	1	0.31	0	0.00	0	0.00	2	0.62	3	0.93	6	1.86
Asp	77	0.84	8	0.52	8	0.52	12	0.78	9	0.59	26	1.70	14	0.91
Glu	93	0.90	11	0.64	25	1.46	1	0.06	12	0.70	17	0.99	27	1.58
Phe	49	0.78	8	0.76	12	1.15	7	0.67	1	0.10	6	0.57	15	1.43
Gly	247	2.12	15	0.77	6	0.31	171	8.81	28	1.44	4	0.21	23	1.19
His	55	1.50	8	1.31	5	0.82	12	1.96	7	1.14	13	2.12	10	1.63
Ile	34	0.38	9	0.61	12	0.81	0	0.00	1	0.07	8	0.54	4	0.27
Lys	85	1.00	18	1.27	27	1.90	4	0.28	6	0.42	23	1.62	7	0.49
Leu	95	0.66	35	1.47	14	0.59	2	0.08	6	0.25	13	0.54	25	1.05
Met	10	0.38	2	0.45	2	0.45	1	0.23	1	0.23	3	0.68	1	0.23
Asn	61	0.95	10	0.93	5	0.47	24	2.23	1	0.09	8	0.74	13	1.21
Pro	114	1.58	22	1.83	45	3.75	0	0.00	0	0.00	46	3.83	1	0.08
Gln	41	0.72	10	1.06	9	0.95	2	0.21	6	0.63	7	0.74	7	0.74
Arg	62	0.78	11	0.83	12	0.91	8	0.61	10	0.76	12	0.91	9	0.68
Ser	125	1.39	9	0.60	14	0.93	3	0.20	68	4.52	15	1.00	16	1.06
Thr	145	1.74	20	1.44	11	0.79	2	0.14	79	5.68	16	1.15	17	1.22
Val	58	0.52	22	1.19	12	0.65	0	0.00	1	0.05	10	0.54	13	0.71
Trp	17	0.78	5	1.37	5	1.37	1	0.27	1	0.27	2	0.55	3	0.82
Tyr	50	0.92	15	1.66	9	1.00	0	0.00	6	0.66	2	0.22	18	1.99
Other	6	0.78	0	0.00	2	1.57	0	0.00	2	1.57	1	0.78	1	0.78

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-31														
Ala	92	1.13	13	0.96	21	1.55	3	0.22	18	1.33	25	1.84	12	0.88
Cys	15	1.26	7	3.54	0	0.00	0	0.00	6	3.03	1	0.51	1	0.51
Asp	52	0.92	7	0.75	9	0.96	4	0.43	11	1.17	14	1.49	7	0.75
Glu	49	0.78	6	0.57	0	0.00	0	0.00	15	1.43	15	1.43	13	1.24
Phe	25	0.65	4	0.62	12	1.87	0	0.00	1	0.16	5	0.78	3	0.47
Gly	240	3.37	14	1.18	18	1.51	134	11.28	10	0.84	4	0.34	60	5.05
His	17	0.75	3	0.80	2	0.53	3	0.80	1	0.27	7	1.86	1	0.27
Ile	25	0.46	7	0.77	3	0.33	0	0.00	11	1.21	3	0.33	1	0.11
Lys	32	0.61	6	0.69	7	0.80	2	0.23	10	1.15	6	0.69	1	0.11
Leu	62	0.71	15	1.03	26	1.78	1	0.07	8	0.55	10	0.68	2	0.14
Met	13	0.80	1	0.37	6	2.21	0	0.00	3	1.10	2	0.74	1	0.37
Asn	35	0.89	15	2.28	2	0.30	1	0.15	5	0.76	6	0.91	6	0.91
Pro	26	0.59	8	1.09	3	0.41	0	0.00	13	1.77	1	0.14	1	0.14
Gln	26	0.75	6	1.03	5	0.86	0	0.00	7	1.21	5	0.86	3	0.52
Arg	27	0.56	9	1.11	5	0.62	1	0.12	4	0.50	7	0.87	1	0.12
Ser	88	1.59	14	1.52	10	1.09	6	0.65	15	1.63	15	1.63	28	3.04
Thr	49	0.96	7	0.82	7	0.82	2	0.23	7	0.82	10	1.17	16	1.88
Val	19	0.28	5	0.44	1	0.09	0	0.00	9	0.80	4	0.35	0	0.00
Trp	12	0.89	1	0.45	4	1.79	0	0.00	1	0.45	6	2.68	0	0.00
Tyr	41	1.23	10	1.81	15	2.71	1	0.18	2	0.36	12	2.17	1	0.18
Other	3	0.64	0	0.00	2	2.56	0	0.00	1	1.28	0	0.00	0	0.00
type: o6-32														
Ala	62	0.92	10	0.89	7	0.62	19	1.69	1	0.09	12	1.07	13	1.16
Cys	9	0.91	2	1.22	2	1.22	1	0.61	0	0.00	2	1.22	2	1.22
Asp	41	0.88	4	0.51	10	1.28	4	0.51	0	0.00	13	1.67	10	1.28
Glu	46	0.88	5	0.57	8	0.92	17	1.95	0	0.00	15	1.72	1	0.11
Phe	26	0.81	2	0.38	6	1.13	5	0.94	0	0.00	10	1.88	3	0.56
Gly	53	0.90	10	1.02	3	0.30	5	0.51	0	0.00	8	0.81	27	2.74
His	22	1.18	5	1.61	4	1.29	0	0.00	1	0.32	4	1.29	8	2.57
Ile	20	0.44	7	0.93	4	0.53	4	0.53	0	0.00	2	0.27	3	0.40
Lys	27	0.62	4	0.55	6	0.83	4	0.55	1	0.14	6	0.83	6	0.83
Leu	49	0.67	12	0.99	14	1.16	6	0.50	0	0.00	12	0.99	5	0.41
Met	8	0.59	4	1.78	1	0.44	1	0.44	0	0.00	0	0.00	2	0.89
Asn	31	0.95	4	0.73	7	1.28	7	1.28	0	0.00	3	0.55	10	1.83
Pro	171	4.67	13	2.13	15	2.46	18	2.95	122	20.00	1	0.16	2	0.33
Gln	20	0.69	3	0.62	3	0.62	4	0.83	0	0.00	8	1.66	2	0.42
Arg	34	0.85	11	1.64	8	1.19	2	0.30	0	0.00	4	0.60	9	1.34
Ser	38	0.83	4	0.52	10	1.31	7	0.92	2	0.26	5	0.65	10	1.31
Thr	32	0.75	6	0.85	10	1.42	0	0.00	3	0.42	10	1.42	3	0.42
Val	40	0.71	13	1.39	10	1.07	5	0.53	1	0.11	3	0.32	8	0.85
Trp	14	1.26	3	1.62	0	0.00	3	1.62	0	0.00	5	2.70	3	1.62
Tyr	40	1.45	8	1.74	3	0.65	17	3.70	0	0.00	8	1.74	4	0.87
Other	3	0.77	1	1.54	0	0.00	2	3.09	0	0.00	0	0.00	0	0.00

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-33														
Ala	49	0.73	7	0.63	6	0.54	16	1.43	4	0.36	4	0.36	12	1.07
Cys	5	0.51	1	0.61	2	1.23	1	0.61	0	0.00	0	0.00	1	0.61
Asp	49	1.06	5	0.65	4	0.52	12	1.55	13	1.68	10	1.29	5	0.65
Glu	40	0.77	11	1.27	3	0.35	15	1.74	6	0.69	2	0.23	3	0.35
Phe	29	0.92	7	1.33	6	1.14	1	0.19	7	1.33	5	0.95	3	0.57
Gly	166	2.83	6	0.61	33	3.38	5	0.51	47	4.81	61	6.24	14	1.43
His	19	1.03	4	1.30	3	0.97	3	0.97	6	1.94	1	0.32	2	0.65
Ile	17	0.38	5	0.67	5	0.67	1	0.13	0	0.00	0	0.00	6	0.80
Lys	34	0.79	9	1.26	2	0.28	8	1.12	4	0.56	0	0.00	11	1.53
Leu	61	0.85	12	1.00	31	2.58	4	0.33	2	0.17	2	0.17	10	0.83
Met	8	0.60	0	0.00	2	0.90	1	0.45	1	0.45	3	1.34	1	0.45
Asn	54	1.66	2	0.37	3	0.55	2	0.37	22	4.06	19	3.51	6	1.11
Pro	48	1.32	8	1.32	4	0.66	36	5.95	0	0.00	0	0.00	0	0.00
Gln	32	1.12	8	1.68	6	1.26	1	0.21	5	1.05	3	0.63	9	1.89
Arg	38	0.95	7	1.05	2	0.30	3	0.45	3	0.45	5	0.75	18	2.71
Ser	33	0.73	5	0.66	1	0.13	11	1.45	5	0.66	4	0.53	7	0.92
Thr	32	0.76	12	1.71	8	1.14	4	0.57	3	0.43	0	0.00	5	0.71
Val	27	0.48	10	1.08	4	0.43	2	0.22	0	0.00	0	0.00	11	1.18
Trp	10	0.91	5	2.72	1	0.54	0	0.00	0	0.00	3	1.63	1	0.54
Tyr	28	1.02	6	1.32	3	0.66	4	0.88	2	0.44	8	1.76	5	1.10
Other	1	0.26	0	0.00	1	1.56	0	0.00	0	0.00	0	0.00	0	0.00
type: o6-34														
Ala	38	0.73	13	1.50	12	1.38	5	0.58	1	0.12	2	0.23	5	0.58
Cys	2	0.26	1	0.79	0	0.00	0	0.00	0	0.00	0	0.00	1	0.79
Asp	68	1.89	6	1.00	7	1.17	38	6.33	10	1.67	0	0.00	7	1.17
Glu	30	0.75	11	1.64	12	1.79	0	0.00	3	0.45	0	0.00	4	0.60
Phe	17	0.69	1	0.24	3	0.73	4	0.98	3	0.73	4	0.98	2	0.49
Gly	126	2.76	2	0.26	2	0.26	3	0.39	47	6.19	69	9.08	3	0.39
His	25	1.74	1	0.42	2	0.83	8	3.33	8	3.33	3	1.25	3	1.25
Ile	11	0.32	3	0.52	2	0.35	2	0.35	0	0.00	0	0.00	4	0.69
Lys	38	1.14	5	0.90	6	1.08	2	0.36	5	0.90	1	0.18	19	3.41
Leu	43	0.77	8	0.86	10	1.07	6	0.64	4	0.43	3	0.32	12	1.28
Met	1	0.10	0	0.00	0	0.00	0	0.00	1	0.58	0	0.00	0	0.00
Asn	41	1.62	0	0.00	5	1.19	10	2.38	11	2.61	9	2.14	6	1.43
Pro	3	0.11	2	0.43	1	0.21	0	0.00	0	0.00	0	0.00	0	0.00
Gln	10	0.45	4	1.08	0	0.00	2	0.54	1	0.27	0	0.00	3	0.81
Arg	26	0.84	3	0.58	6	1.16	3	0.58	2	0.39	1	0.19	11	2.13
Ser	55	1.56	30	5.10	13	2.21	5	0.85	2	0.34	1	0.17	4	0.68
Thr	30	0.92	6	1.10	10	1.84	8	1.47	2	0.37	0	0.00	4	0.73
Val	18	0.42	2	0.28	6	0.83	3	0.42	0	0.00	0	0.00	7	0.97
Trp	6	0.70	0	0.00	1	0.70	1	0.70	0	0.00	3	2.10	1	0.70
Tyr	18	0.85	3	0.85	3	0.85	1	0.28	1	0.28	5	1.41	5	1.41
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-35														
Ala	19	0.32	1	0.10	4	0.40	3	0.30	7	0.71	0	0.00	4	0.40
Cys	7	0.81	1	0.69	2	1.39	2	1.39	0	0.00	0	0.00	2	1.39
Asp	44	1.08	2	0.29	7	1.02	1	0.15	31	4.53	2	0.29	1	0.15
Glu	36	0.79	7	0.92	8	1.05	6	0.79	9	1.18	0	0.00	6	0.79
Phe	25	0.89	8	1.71	5	1.07	7	1.50	3	0.64	0	0.00	2	0.43
Gly	120	2.32	6	0.69	0	0.00	0	0.00	8	0.93	103	11.91	3	0.35
His	22	1.35	7	2.56	2	0.73	5	1.83	6	2.20	0	0.00	2	0.73
Ile	40	1.01	14	2.12	6	0.91	13	1.97	0	0.00	0	0.00	7	1.06
Lys	57	1.50	8	1.26	10	1.58	11	1.73	6	0.95	1	0.16	21	3.31
Leu	51	0.80	15	1.41	18	1.69	11	1.03	1	0.09	0	0.00	6	0.56
Met	6	0.51	2	1.01	0	0.00	0	0.00	1	0.51	1	0.51	2	1.01
Asn	55	1.92	4	0.83	12	2.50	3	0.63	26	5.43	4	0.83	6	1.25
Pro	9	0.28	4	0.75	5	0.93	0	0.00	0	0.00	0	0.00	0	0.00
Gln	25	0.99	2	0.47	2	0.47	3	0.71	5	1.18	0	0.00	13	3.08
Arg	22	0.63	3	0.51	4	0.68	3	0.51	2	0.34	0	0.00	10	1.70
Ser	21	0.52	4	0.60	2	0.30	4	0.60	6	0.90	2	0.30	3	0.45
Thr	22	0.59	5	0.81	9	1.45	3	0.48	0	0.00	0	0.00	5	0.81
Val	69	1.40	17	2.07	9	1.10	31	3.77	0	0.00	0	0.00	12	1.46
Trp	9	0.92	1	0.61	3	1.84	2	1.23	1	0.61	0	0.00	2	1.23
Tyr	27	1.12	4	0.99	5	1.24	7	1.74	3	0.74	2	0.50	6	1.49
Other	4	1.18	0	0.00	2	3.52	0	0.00	0	0.00	0	0.00	2	3.52
type: o6-36														
Ala	37	0.82	11	1.46	7	0.93	10	1.32	2	0.26	1	0.13	6	0.79
Cys	13	1.97	4	3.63	1	0.91	5	4.54	1	0.91	0	0.00	2	1.82
Asp	15	0.48	1	0.19	6	1.15	3	0.57	5	0.96	0	0.00	0	0.00
Glu	20	0.57	2	0.34	4	0.68	4	0.68	7	1.20	0	0.00	3	0.51
Phe	16	0.75	1	0.28	5	1.40	3	0.84	6	1.68	0	0.00	1	0.28
Gly	25	0.63	8	1.21	4	0.60	9	1.36	0	0.00	0	0.00	4	0.60
His	13	1.04	3	1.43	4	1.91	3	1.43	2	0.96	0	0.00	1	0.48
Ile	27	0.89	9	1.78	1	0.20	3	0.59	6	1.19	0	0.00	8	1.58
Lys	20	0.69	1	0.21	5	1.03	4	0.82	5	1.03	0	0.00	5	1.03
Leu	31	0.64	8	0.98	9	1.11	0	0.00	5	0.61	1	0.12	8	0.98
Met	10	1.10	2	1.32	1	0.66	2	1.32	2	1.32	0	0.00	3	1.98
Asn	14	0.64	4	1.09	2	0.55	3	0.82	2	0.55	0	0.00	3	0.82
Pro	113	4.60	10	2.44	6	1.46	0	0.00	4	0.98	86	20.99	7	1.71
Gln	16	0.83	2	0.62	4	1.24	3	0.93	3	0.93	0	0.00	4	1.24
Arg	16	0.59	2	0.44	4	0.89	1	0.22	5	1.11	0	0.00	4	0.89
Ser	27	0.88	3	0.59	8	1.56	7	1.37	7	1.37	0	0.00	2	0.39
Thr	54	1.90	4	0.84	10	2.11	15	3.16	11	2.32	0	0.00	14	2.95
Val	37	0.98	10	1.59	2	0.32	9	1.43	10	1.59	0	0.00	6	0.95
Trp	6	0.80	2	1.61	0	0.00	1	0.80	2	1.61	0	0.00	1	0.80
Tyr	16	0.87	1	0.32	4	1.30	3	0.97	3	0.97	0	0.00	5	1.62
Other	2	0.77	0	0.00	1	2.30	0	0.00	0	0.00	0	0.00	1	2.30

Table A3.13 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of open 6 turns. Only turn types with at least 100 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***		***position 6***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: o6-37														
Ala	25	0.43	6	0.62	5	0.52	6	0.62	0	0.00	3	0.31	5	0.52
Cys	7	0.82	3	2.12	0	0.00	0	0.00	0	0.00	2	1.41	2	1.41
Asp	41	1.02	4	0.60	0	0.00	3	0.45	5	0.74	3	0.45	26	3.87
Glu	39	0.87	5	0.67	6	0.80	14	1.86	2	0.27	9	1.20	3	0.40
Phe	17	0.62	7	1.53	2	0.44	0	0.00	1	0.22	4	0.87	3	0.65
Gly	129	2.53	13	1.53	0	0.00	2	0.24	86	10.12	2	0.24	26	3.06
His	13	0.81	0	0.00	2	0.75	3	1.12	0	0.00	4	1.49	4	1.49
Ile	49	1.26	12	1.85	15	2.31	3	0.46	0	0.00	17	2.62	2	0.31
Lys	38	1.02	4	0.64	11	1.77	8	1.28	1	0.16	13	2.09	1	0.16
Leu	42	0.67	12	1.15	10	0.96	11	1.05	1	0.10	4	0.38	4	0.38
Met	11	0.94	2	1.03	3	1.54	2	1.03	1	0.51	1	0.51	2	1.03
Asn	27	0.96	0	0.00	4	0.85	2	0.42	12	2.55	0	0.00	9	1.91
Pro	38	1.20	6	1.14	1	0.19	31	5.89	0	0.00	0	0.00	0	0.00
Gln	28	1.12	5	1.21	11	2.65	1	0.24	1	0.24	7	1.69	3	0.72
Arg	27	0.78	3	0.52	9	1.56	8	1.38	0	0.00	6	1.04	1	0.17
Ser	22	0.56	2	0.30	4	0.61	2	0.30	1	0.15	3	0.46	10	1.52
Thr	27	0.74	4	0.66	7	1.15	3	0.49	0	0.00	4	0.66	9	1.48
Val	64	1.32	9	1.11	21	2.60	7	0.87	0	0.00	27	3.34	0	0.00
Trp	12	1.25	6	3.75	1	0.63	3	1.88	1	0.63	1	0.63	0	0.00
Tyr	22	0.93	10	2.52	1	0.25	4	1.01	1	0.25	3	0.76	3	0.76
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
type: o6-38														
Ala	55	1.02	10	1.11	6	0.67	12	1.33	8	0.89	1	0.11	18	2.00
Cys	12	1.52	2	1.52	4	3.04	2	1.52	1	0.76	0	0.00	3	2.28
Asp	23	0.61	3	0.48	11	1.76	1	0.16	4	0.64	3	0.48	1	0.16
Glu	28	0.67	6	0.86	3	0.43	9	1.29	7	1.00	1	0.14	2	0.29
Phe	20	0.78	8	1.88	5	1.17	1	0.23	3	0.70	3	0.70	0	0.00
Gly	148	3.12	13	1.65	2	0.25	30	3.80	2	0.25	85	10.76	16	2.03
His	8	0.53	0	0.00	3	1.20	1	0.40	0	0.00	1	0.40	3	1.20
Ile	30	0.83	4	0.66	16	2.66	7	1.16	1	0.17	0	0.00	2	0.33
Lys	33	0.95	5	0.86	3	0.52	11	1.90	8	1.38	0	0.00	6	1.04
Leu	42	0.72	14	1.44	13	1.34	5	0.51	4	0.41	0	0.00	6	0.62
Met	9	0.83	2	1.11	2	1.11	0	0.00	2	1.11	0	0.00	3	1.66
Asn	22	0.84	4	0.91	1	0.23	1	0.23	3	0.69	6	1.37	7	1.60
Pro	38	1.30	7	1.43	1	0.20	0	0.00	29	5.93	0	0.00	1	0.20
Gln	14	0.61	2	0.52	2	0.52	1	0.26	6	1.56	2	0.52	1	0.26
Arg	24	0.75	1	0.19	7	1.30	8	1.49	4	0.75	0	0.00	4	0.75
Ser	19	0.52	7	1.14	2	0.33	1	0.16	4	0.65	1	0.16	4	0.65
Thr	28	0.82	5	0.88	5	0.88	6	1.06	3	0.53	0	0.00	9	1.59
Val	48	1.07	5	0.67	13	1.73	6	0.80	14	1.87	0	0.00	10	1.33
Trp	4	0.45	0	0.00	0	0.00	0	0.00	1	0.67	2	1.35	1	0.67
Tyr	22	1.00	5	1.36	5	1.36	3	0.82	1	0.27	0	0.00	8	2.17
Other	3	0.96	2	3.85	1	1.93	0	0.00	0	0.00	0	0.00	0	0.00

REVERSE 2 TURNS

Table A3.14. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 2 turns. Only turn types with at least seven representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*	
	n	***P***	***n***	***P***	***n***	***P***
type: r2-1						
Ala	52	0.38	21	0.31	31	0.46
Cys	2	0.10	0	0.00	2	0.19
Asp	18	0.19	9	0.19	9	0.19
Glu	22	0.20	8	0.15	14	0.26
Phe	10	0.15	1	0.03	9	0.27
Gly	121	1.03	32	0.54	89	1.51
His	7	0.18	1	0.05	6	0.32
Ile	5	0.05	3	0.06	2	0.04
Lys	25	0.27	12	0.26	13	0.28
Leu	59	0.39	50	0.65	9	0.12
Met	3	0.11	2	0.14	1	0.07
Asn	99	1.44	94	2.73	5	0.15
Pro	4	0.05	0	0.00	4	0.11
Gln	12	0.20	9	0.30	3	0.10
Arg	20	0.24	8	0.19	12	0.29
Ser	16	0.17	11	0.23	5	0.10
Thr	14	0.16	3	0.07	11	0.25
Val	7	0.06	6	0.10	1	0.02
Trp	530	23.78	528	47.39	2	0.18
Tyr	594	10.43	12	0.42	582	20.43
Other	0	0.00	0	0.00	0	0.00
type: r2-2						
Ala	18	1.37	18	2.73	0	0.00
Cys	0	0.00	0	0.00	0	0.00
Asp	6	0.64	5	1.07	1	0.21
Glu	3	0.28	3	0.56	0	0.00
Phe	7	1.09	7	2.17	0	0.00
Gly	77	6.69	3	0.52	74	12.87
His	4	1.08	4	2.16	0	0.00
Ile	0	0.00	0	0.00	0	0.00
Lys	1	0.11	0	0.00	1	0.22
Leu	2	0.13	1	0.13	1	0.13
Met	2	0.72	2	1.44	0	0.00
Asn	3	0.45	1	0.30	2	0.60
Pro	0	0.00	0	0.00	0	0.00
Gln	1	0.17	1	0.35	0	0.00
Arg	1	0.12	1	0.25	0	0.00
Ser	29	3.10	29	6.21	0	0.00
Thr	3	0.35	3	0.70	0	0.00
Val	0	0.00	0	0.00	0	0.00
Trp	0	0.00	0	0.00	0	0.00
Tyr	0	0.00	0	0.00	0	0.00
Other	1	1.37	1	2.73	0	0.00

Table A3.14 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 2 turns. Only turn types with at least seven representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***	
	n	***P***	***n***	***P***	***n***	***P***
type: r2-3						
Ala	1	0.23	1	0.44	0	0.00
Cys	3	4.46	0	0.00	3	8.76
Asp	1	0.32	1	0.63	0	0.00
Glu	4	1.12	3	1.65	1	0.55
Phe	3	1.39	2	1.82	1	0.91
Gly	4	1.04	1	0.51	3	1.53
His	1	0.80	1	1.58	0	0.00
Ile	0	0.00	0	0.00	0	0.00
Lys	2	0.67	1	0.65	1	0.65
Leu	2	0.40	0	0.00	2	0.78
Met	0	0.00	0	0.00	0	0.00
Asn	4	1.78	4	3.49	0	0.00
Pro	11	4.57	0	0.00	11	8.97
Gln	2	1.03	0	0.00	2	2.03
Arg	0	0.00	0	0.00	0	0.00
Ser	7	2.23	7	4.38	0	0.00
Thr	7	2.44	6	4.11	1	0.69
Val	0	0.00	0	0.00	0	0.00
Trp	1	1.37	0	0.00	1	2.69
Tyr	0	0.00	0	0.00	0	0.00
Other	1	4.07	0	0.00	1	7.99
type: r2-4						
Ala	5	1.30	3	1.56	2	1.04
Cys	0	0.00	0	0.00	0	0.00
Asp	3	1.11	2	1.48	1	0.74
Glu	8	2.58	3	1.94	5	3.23
Phe	1	0.53	0	0.00	1	1.07
Gly	9	2.69	5	2.99	4	2.39
His	1	0.93	1	1.85	0	0.00
Ile	3	1.12	2	1.49	1	0.74
Lys	1	0.38	1	0.77	0	0.00
Leu	1	0.23	0	0.00	1	0.46
Met	0	0.00	0	0.00	0	0.00
Asn	4	2.05	1	1.02	3	3.07
Pro	0	0.00	0	0.00	0	0.00
Gln	3	1.78	1	1.19	2	2.38
Arg	2	0.85	1	0.85	1	0.85
Ser	1	0.37	1	0.74	0	0.00
Thr	2	0.80	1	0.80	1	0.80
Val	2	0.61	1	0.61	1	0.61
Trp	0	0.00	0	0.00	0	0.00
Tyr	0	0.00	0	0.00	0	0.00
Other	0	0.00	0	0.00	0	0.00

Table A3.14 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 2 turns. Only turn types with at least seven representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***	
	n	***P***	***n***	***P***	***n***	***P***
type: r2-5						
Ala	1	0.75	1	1.50	0	0.00
Cys	0	0.00	0	0.00	0	0.00
Asp	1	1.06	1	2.12	0	0.00
Glu	1	0.93	0	0.00	1	1.86
Phe	0	0.00	0	0.00	0	0.00
Gly	7	6.01	6	10.30	1	1.72
His	0	0.00	0	0.00	0	0.00
Ile	0	0.00	0	0.00	0	0.00
Lys	0	0.00	0	0.00	0	0.00
Leu	0	0.00	0	0.00	0	0.00
Met	0	0.00	0	0.00	0	0.00
Asn	0	0.00	0	0.00	0	0.00
Pro	5	6.88	0	0.00	5	13.76
Gln	0	0.00	0	0.00	0	0.00
Arg	0	0.00	0	0.00	0	0.00
Ser	0	0.00	0	0.00	0	0.00
Thr	0	0.00	0	0.00	0	0.00
Val	0	0.00	0	0.00	0	0.00
Trp	0	0.00	0	0.00	0	0.00
Tyr	1	1.78	0	0.00	1	3.55
Other	0	0.00	0	0.00	0	0.00

REVERSE 3 TURNS

Table A3.15. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 3 turns. Only turn types with at least seven representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r3-1								
Ala	11	0.49	4	0.53	1	0.13	6	0.79
Cys	2	0.61	1	0.91	0	0.00	1	0.91
Asp	17	1.08	6	1.15	8	1.53	3	0.57
Glu	13	0.74	0	0.00	8	1.37	5	0.86
Phe	5	0.47	0	0.00	2	0.56	3	0.84
Gly	67	3.37	53	8.01	2	0.30	12	1.81
His	12	1.91	0	0.00	10	4.78	2	0.96
Ile	7	0.46	0	0.00	4	0.79	3	0.59
Lys	15	1.03	4	0.82	7	1.44	4	0.82
Leu	12	0.49	1	0.12	7	0.86	4	0.49
Met	2	0.44	1	0.66	1	0.66	0	0.00
Asn	8	0.73	0	0.00	5	1.36	3	0.82
Pro	1	0.08	0	0.00	1	0.24	0	0.00
Gln	12	1.24	2	0.62	3	0.93	7	2.17
Arg	6	0.44	0	0.00	4	0.89	2	0.44
Ser	22	1.43	9	1.76	7	1.37	6	1.17
Thr	27	1.90	5	1.05	10	2.11	12	2.53
Val	8	0.42	0	0.00	4	0.64	4	0.64
Trp	6	1.61	2	1.61	1	0.80	3	2.41
Tyr	11	1.19	0	0.00	3	0.97	8	2.59
Other	0	0.00	0	0.00	0	0.00	0	0.00
type: r3-2								
Ala	0	0.00	0	0.00	0	0.00	0	0.00
Cys	0	0.00	0	0.00	0	0.00	0	0.00
Asp	4	3.20	3	7.21	0	0.00	1	2.40
Glu	2	1.43	0	0.00	0	0.00	2	4.30
Phe	0	0.00	0	0.00	0	0.00	0	0.00
Gly	10	6.33	2	3.80	7	13.30	1	1.90
His	0	0.00	0	0.00	0	0.00	0	0.00
Ile	0	0.00	0	0.00	0	0.00	0	0.00
Lys	0	0.00	0	0.00	0	0.00	0	0.00
Leu	0	0.00	0	0.00	0	0.00	0	0.00
Met	1	2.77	1	8.31	0	0.00	0	0.00
Asn	2	2.29	0	0.00	0	0.00	2	6.86
Pro	0	0.00	0	0.00	0	0.00	0	0.00
Gln	0	0.00	0	0.00	0	0.00	0	0.00
Arg	0	0.00	0	0.00	0	0.00	0	0.00
Ser	1	0.82	1	2.45	0	0.00	0	0.00
Thr	0	0.00	0	0.00	0	0.00	0	0.00
Val	0	0.00	0	0.00	0	0.00	0	0.00
Trp	0	0.00	0	0.00	0	0.00	0	0.00
Tyr	1	1.36	0	0.00	0	0.00	1	4.08
Other	0	0.00	0	0.00	0	0.00	0	0.00

Table A3.15 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 3 turns. Only turn types with at least seven representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r3-3								
Ala	1	0.49	0	0.00	1	1.46	0	0.00
Cys	0	0.00	0	0.00	0	0.00	0	0.00
Asp	1	0.70	0	0.00	0	0.00	1	2.10
Glu	1	0.63	0	0.00	1	1.88	0	0.00
Phe	2	2.05	1	3.08	1	3.08	0	0.00
Gly	2	1.11	0	0.00	2	3.32	0	0.00
His	0	0.00	0	0.00	0	0.00	0	0.00
Ile	0	0.00	0	0.00	0	0.00	0	0.00
Lys	2	1.51	0	0.00	0	0.00	2	4.53
Leu	0	0.00	0	0.00	0	0.00	0	0.00
Met	0	0.00	0	0.00	0	0.00	0	0.00
Asn	4	4.00	3	9.00	1	3.00	0	0.00
Pro	0	0.00	0	0.00	0	0.00	0	0.00
Gln	3	3.40	1	3.40	0	0.00	2	6.81
Arg	1	0.82	0	0.00	1	2.45	0	0.00
Ser	1	0.72	1	2.15	0	0.00	0	0.00
Thr	2	1.54	0	0.00	1	2.32	1	2.32
Val	1	0.58	0	0.00	0	0.00	1	1.75
Trp	0	0.00	0	0.00	0	0.00	0	0.00
Tyr	3	3.57	2	7.13	0	0.00	1	3.57
Other	0	0.00	0	0.00	0	0.00	0	0.00
type: r3-4								
Ala	0	0.00	0	0.00	0	0.00	0	0.00
Cys	0	0.00	0	0.00	0	0.00	0	0.00
Asp	7	4.91	4	8.41	3	6.31	0	0.00
Glu	1	0.63	0	0.00	1	1.88	0	0.00
Phe	0	0.00	0	0.00	0	0.00	0	0.00
Gly	8	4.43	0	0.00	0	0.00	8	13.30
His	0	0.00	0	0.00	0	0.00	0	0.00
Ile	0	0.00	0	0.00	0	0.00	0	0.00
Lys	0	0.00	0	0.00	0	0.00	0	0.00
Leu	0	0.00	0	0.00	0	0.00	0	0.00
Met	0	0.00	0	0.00	0	0.00	0	0.00
Asn	1	1.00	0	0.00	1	3.00	0	0.00
Pro	0	0.00	0	0.00	0	0.00	0	0.00
Gln	0	0.00	0	0.00	0	0.00	0	0.00
Arg	1	0.82	0	0.00	1	2.45	0	0.00
Ser	0	0.00	0	0.00	0	0.00	0	0.00
Thr	0	0.00	0	0.00	0	0.00	0	0.00
Val	0	0.00	0	0.00	0	0.00	0	0.00
Trp	1	2.94	1	8.83	0	0.00	0	0.00
Tyr	4	4.76	3	10.70	1	3.57	0	0.00
Other	1	8.43	0	0.00	1	25.29	0	0.00

Table A3.15 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 3 turns. Only turn types with at least seven representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r3-5								
Ala	2	1.11	1	1.66	0	0.00	1	1.66
Cys	0	0.00	0	0.00	0	0.00	0	0.00
Asp	2	1.60	1	2.40	0	0.00	1	2.40
Glu	1	0.72	0	0.00	0	0.00	1	2.15
Phe	0	0.00	0	0.00	0	0.00	0	0.00
Gly	1	0.63	1	1.90	0	0.00	0	0.00
His	1	2.00	0	0.00	0	0.00	1	6.01
Ile	0	0.00	0	0.00	0	0.00	0	0.00
Lys	0	0.00	0	0.00	0	0.00	0	0.00
Leu	2	1.03	1	1.54	0	0.00	1	1.54
Met	0	0.00	0	0.00	0	0.00	0	0.00
Asn	1	1.14	1	3.43	0	0.00	0	0.00
Pro	7	7.16	0	0.00	7	21.48	0	0.00
Gln	0	0.00	0	0.00	0	0.00	0	0.00
Arg	1	0.93	0	0.00	0	0.00	1	2.79
Ser	1	0.82	1	2.45	0	0.00	0	0.00
Thr	1	0.88	1	2.65	0	0.00	0	0.00
Val	0	0.00	0	0.00	0	0.00	0	0.00
Trp	0	0.00	0	0.00	0	0.00	0	0.00
Tyr	1	1.36	0	0.00	0	0.00	1	4.08
Other	0	0.00	0	0.00	0	0.00	0	0.00

REVERSE 4 TURNS

Table A3.16. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 4 turns. Only turn types with at least 50 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*		*position 4*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r4-1										
Ala	228	0.40	74	0.52	77	0.54	9	0.06	68	0.48
Cys	46	0.56	21	1.01	14	0.68	2	0.10	9	0.43
Asp	541	1.38	43	0.44	406	4.13	68	0.69	24	0.24
Glu	442	1.01	126	1.15	115	1.05	25	0.23	176	1.60
Phe	180	0.67	103	1.53	33	0.49	12	0.18	32	0.48
Gly	1,497	3.01	24	0.19	156	1.25	1,298	10.43	19	0.15
His	187	1.19	63	1.60	59	1.50	17	0.43	48	1.22
Ile	283	0.75	191	2.01	0	0.00	0	0.00	92	0.97
Lys	606	1.66	138	1.51	89	0.98	20	0.22	359	3.94
Leu	233	0.38	125	0.82	26	0.17	5	0.03	77	0.50
Met	48	0.42	13	0.46	10	0.35	3	0.11	22	0.77
Asn	646	2.35	25	0.36	465	6.75	106	1.54	50	0.73
Pro	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Gln	271	1.12	60	0.99	59	0.97	14	0.23	138	2.27
Arg	361	1.07	116	1.37	52	0.61	24	0.28	169	2.00
Ser	216	0.56	75	0.78	48	0.50	36	0.37	57	0.59
Thr	111	0.31	35	0.39	0	0.00	2	0.02	74	0.83
Val	446	0.94	275	2.33	3	0.03	1	0.01	167	1.41
Trp	46	0.49	30	1.28	5	0.21	0	0.00	11	0.47
Tyr	209	0.90	114	1.97	35	0.60	10	0.17	50	0.86
Other	19	0.58	3	0.37	2	0.24	2	0.24	12	1.47
type: r4-2										
Ala	265	0.64	26	0.25	118	1.14	63	0.61	58	0.56
Cys	34	0.56	12	0.79	2	0.13	6	0.40	14	0.93
Asp	486	1.69	144	2.01	136	1.89	175	2.44	31	0.43
Glu	371	1.16	25	0.31	177	2.21	109	1.36	60	0.75
Phe	126	0.64	23	0.47	21	0.43	35	0.71	47	0.96
Gly	683	1.88	206	2.27	40	0.44	33	0.36	404	4.45
His	147	1.28	19	0.66	17	0.59	52	1.81	59	2.06
Ile	84	0.30	6	0.09	22	0.32	25	0.36	31	0.45
Lys	242	0.91	10	0.15	104	1.56	85	1.28	43	0.65
Leu	165	0.37	33	0.30	38	0.34	47	0.42	47	0.42
Met	35	0.42	6	0.29	6	0.29	11	0.53	12	0.58
Asn	335	1.67	126	2.50	53	1.05	115	2.28	41	0.81
Pro	197	0.88	0	0.00	195	3.47	2	0.04	0	0.00
Gln	135	0.76	12	0.27	28	0.63	58	1.31	37	0.83
Arg	195	0.79	14	0.23	54	0.87	77	1.25	50	0.81
Ser	511	1.82	210	2.98	100	1.42	119	1.69	82	1.16
Thr	512	1.97	300	4.60	50	0.77	103	1.58	59	0.91
Val	114	0.33	12	0.14	27	0.31	18	0.21	57	0.66
Trp	48	0.70	4	0.23	11	0.64	14	0.82	19	1.11
Tyr	136	0.80	19	0.45	9	0.21	57	1.35	51	1.20
Other	11	0.46	1	0.17	0	0.00	4	0.67	6	1.00

Table A3.16 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 4 turns. Only turn types with at least 50 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r4-3										
Ala	123	0.43	39	0.55	15	0.21	39	0.55	30	0.42
Cys	30	0.73	17	1.64	1	0.10	4	0.39	8	0.77
Asp	380	1.94	73	1.48	54	1.10	206	4.19	47	0.96
Glu	248	1.13	75	1.37	26	0.47	77	1.40	70	1.27
Phe	88	0.66	41	1.22	3	0.09	7	0.21	37	1.10
Gly	634	2.55	5	0.08	586	9.42	28	0.45	15	0.24
His	84	1.07	29	1.48	10	0.51	9	0.46	36	1.83
Ile	122	0.64	72	1.52	5	0.11	4	0.08	41	0.86
Lys	205	1.13	54	1.18	30	0.66	47	1.03	74	1.62
Leu	142	0.46	54	0.71	3	0.04	13	0.17	72	0.94
Met	33	0.58	10	0.70	3	0.21	6	0.42	14	0.99
Asn	249	1.81	41	1.19	34	0.99	149	4.32	25	0.73
Pro	34	0.22	0	0.00	0	0.00	34	0.88	0	0.00
Gln	120	0.99	34	1.12	12	0.40	24	0.79	50	1.65
Arg	171	1.01	47	1.11	22	0.52	32	0.76	70	1.66
Ser	190	0.99	53	1.10	12	0.25	81	1.68	44	0.91
Thr	155	0.87	23	0.52	0	0.00	49	1.10	83	1.86
Val	172	0.73	110	1.86	3	0.05	5	0.08	54	0.91
Trp	28	0.60	10	0.85	1	0.09	1	0.09	16	1.37
Tyr	89	0.77	36	1.24	6	0.21	12	0.41	35	1.21
Other	11	0.67	4	0.98	1	0.24	0	0.00	6	1.47
type: r4-4										
Ala	28	0.59	14	1.19	8	0.68	3	0.25	3	0.25
Cys	2	0.29	0	0.00	0	0.00	1	0.58	1	0.58
Asp	37	1.14	6	0.74	19	2.33	9	1.11	3	0.37
Glu	28	0.77	11	1.21	10	1.10	2	0.22	5	0.55
Phe	22	0.99	4	0.72	3	0.54	1	0.18	14	2.52
Gly	142	3.45	30	2.91	0	0.00	90	8.74	22	2.14
His	6	0.46	1	0.31	3	0.92	0	0.00	2	0.61
Ile	20	0.64	4	0.51	7	0.89	0	0.00	9	1.15
Lys	38	1.26	9	1.19	12	1.59	2	0.26	15	1.99
Leu	25	0.49	11	0.87	3	0.24	1	0.08	10	0.79
Met	6	0.64	4	1.70	0	0.00	0	0.00	2	0.85
Asn	32	1.40	1	0.18	8	1.40	15	2.63	8	1.40
Pro	38	1.49	0	0.00	38	5.96	0	0.00	0	0.00
Gln	15	0.75	2	0.40	6	1.19	3	0.60	4	0.80
Arg	24	0.86	7	1.00	5	0.71	3	0.43	9	1.29
Ser	32	1.00	18	2.25	7	0.88	3	0.38	4	0.50
Thr	11	0.37	1	0.14	4	0.54	0	0.00	6	0.81
Val	20	0.51	11	1.12	2	0.20	0	0.00	7	0.71
Trp	5	0.64	1	0.52	0	0.00	1	0.52	3	1.55
Tyr	16	0.83	1	0.21	2	0.42	3	0.62	10	2.08
Other	1	0.37	1	1.48	0	0.00	0	0.00	0	0.00

Table A3.16 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 4 turns. Only turn types with at least 50 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r4-5										
Ala	13	0.40	0	0.00	4	0.50	4	0.50	5	0.62
Cys	5	1.07	0	0.00	1	0.85	0	0.00	4	3.40
Asp	31	1.39	8	1.43	3	0.54	18	3.22	2	0.36
Glu	16	0.64	5	0.80	6	0.96	4	0.64	1	0.16
Phe	17	1.12	2	0.52	1	0.26	1	0.26	13	3.41
Gly	76	2.70	72	10.18	1	0.14	2	0.28	1	0.14
His	7	0.79	0	0.00	1	0.45	6	2.69	0	0.00
Ile	24	1.12	0	0.00	7	1.30	10	1.85	7	1.30
Lys	14	0.68	0	0.00	12	2.31	1	0.19	1	0.19
Leu	37	1.07	0	0.00	3	0.34	23	2.64	11	1.26
Met	1	0.16	0	0.00	0	0.00	0	0.00	1	0.62
Asn	13	0.83	4	1.02	2	0.51	5	1.28	2	0.51
Pro	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Gln	6	0.44	0	0.00	2	0.58	3	0.87	1	0.29
Arg	8	0.42	0	0.00	6	1.25	1	0.21	1	0.21
Ser	3	0.14	1	0.18	0	0.00	1	0.18	1	0.18
Thr	38	1.88	1	0.20	23	4.54	10	1.97	4	0.79
Val	36	1.35	0	0.00	19	2.83	4	0.60	13	1.94
Trp	11	2.08	0	0.00	2	1.50	1	0.75	8	6.01
Tyr	18	1.37	1	0.30	1	0.30	0	0.00	16	4.86
Other	2	1.08	0	0.00	0	0.00	0	0.00	2	4.30
type: r4-6										
Ala	12	0.39	2	0.26	0	0.00	3	0.39	7	0.92
Cys	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Asp	50	2.37	38	7.18	1	0.19	8	1.51	3	0.57
Glu	15	0.64	1	0.17	0	0.00	12	2.03	2	0.34
Phe	6	0.42	3	0.83	0	0.00	1	0.28	2	0.55
Gly	85	3.18	2	0.30	78	11.65	2	0.30	3	0.45
His	12	1.42	3	1.42	0	0.00	7	3.31	2	0.95
Ile	9	0.44	3	0.59	1	0.20	1	0.20	4	0.78
Lys	24	1.23	2	0.41	1	0.20	13	2.65	8	1.63
Leu	10	0.30	2	0.24	1	0.12	3	0.36	4	0.49
Met	4	0.66	1	0.65	1	0.65	0	0.00	2	1.31
Asn	34	2.30	16	4.31	2	0.54	5	1.35	11	2.97
Pro	2	0.12	0	0.00	0	0.00	2	0.48	0	0.00
Gln	8	0.61	0	0.00	0	0.00	4	1.22	4	1.22
Arg	11	0.61	1	0.22	1	0.22	2	0.44	7	1.54
Ser	17	0.82	1	0.19	1	0.19	12	2.31	3	0.58
Thr	25	1.31	10	2.08	0	0.00	6	1.25	9	1.87
Val	15	0.59	2	0.31	1	0.16	4	0.63	8	1.26
Trp	3	0.60	0	0.00	1	0.79	1	0.79	1	0.79
Tyr	13	1.04	2	0.64	0	0.00	3	0.96	8	2.56
Other	1	0.57	0	0.00	0	0.00	0	0.00	1	2.27

REVERSE 5 TURNS

Table A3.17. The means and standard deviations of the dihedral angles (in degrees) and hydorgen bond energies of the classified turn types for the training and test set of reverse 5 turn structures. Turn types r5-23 and r5-30 which are not presented in the manuscript are listed. The means (∅) and standard deviations (±) were calculated using angular statistics. The results for the turns of the test set are given in italics. For a clearer differentiation between the turn types, the colours indicate distinct angle intervals ranging from -180° to 180° in 15° steps according to the colour wheel.

type	|turns|	ψ_i		φ_i		ω_i		ψ_{i+1}		φ_{i+1}		ω_{i+1}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r5-23	7	-143.5	11.4	169.3	3.1	176.7	3.8	59.7	10.6	-132.4	9.0	177.4	3.6
r5-23	*5*	*-139.1*	*10.1*	*167.4*	*3.0*	*175.9*	*2.4*	*60.4*	*7.4*	*-135.1*	*2.6*	*179.2*	*3.1*
r5-24	6	-99.9	18.8	118.1	15.5	-179.6	5.0	-55.6	5.7	140.2	3.9	-0.1	2.0
r5-24	*5*	*-97.1*	*18.2*	*120.4*	*15.0*	*179.5*	*4.1*	*-55.1*	*7.7*	*138.9*	*2.7*	*-2.4*	*4.8*
r5-25	7	-131.9	14.9	118.0	15.1	-170.7	6.8	-58.8	4.8	-30.2	16.8	178.7	3.4
r5-25	*5*	*-141.1*	*10.3*	*112.3*	*15.4*	*-171.7*	*10.7*	*-58.4*	*4.6*	*-26.6*	*6.6*	*179.0*	*2.5*
r5-26	6	-108.9	1.3	10.2	3.2	175.8	2.4	67.5	4.6	-34.6	7.6	-175.6	1.1
r5-26	*5*	*-110.1*	*1.7*	*8.1*	*5.1*	*174.9*	*5.7*	*69.7*	*9.9*	*-33.4*	*10.4*	*-177.3*	*2.6*
r5-27	7	-105.6	13.1	-165.3	16.7	175.5	4.2	-66.5	15.8	135.8	17.7	-178.4	6.5
r5-27	*6*	*-107.4*	*12.0*	*-161.0*	*15.2*	*178.3*	*5.4*	*-69.6*	*19.6*	*148.9*	*12.9*	*178.9*	*6.3*
r5-28	6	101.5	13.0	-5.1	16.3	176.6	2.7	-86.3	19.4	170.0	11.6	178.6	1.5
r5-28	*6*	*98.8*	*12.5*	*-6.1*	*10.1*	*176.2*	*2.4*	*-86.5*	*15.6*	*170.7*	*9.4*	*177.6*	*3.8*
r5-29	5	-126.7	18.9	154.0	25.7	-178.3	4.5	-78.1	19.4	139.4	13.0	-4.8	5.4
r5-29	*3*	*-128.2*	*28.8*	*163.8*	*12.8*	*174.8*	*11.6*	*-72.9*	*24.5*	*134.9*	*10.7*	*-1.8*	*1.3*
r5-30	5	-164.0	25.4	-168.7	22.0	-175.6	5.7	-55.6	5.6	140.6	9.6	-176.3	5.3
r5-30	*4*	*-160.1*	*30.6*	*-165.3*	*23.2*	*-178.5*	*7.1*	*-51.8*	*3.6*	*136.7*	*6.4*	*-177.7*	*2.0*

type	|turns|	ψ_{i+2}		φ_{i+2}		ω_{i+2}		ψ_{i+3}		φ_{i+3}		ω_{i+3}	
		∅	±	∅	±	∅	±	∅	±	∅	±	∅	±
r5-23	7	-94.2	12.0	69.7	16.5	-180.0	2.4	63.0	9.9	33.3	15.4	178.5	2.8
r5-23	*5*	*-95.6*	*7.5*	*72.1*	*4.9*	*176.3*	*4.2*	*65.5*	*8.5*	*35.8*	*8.0*	*175.3*	*5.2*
r5-24	6	-88.0	10.5	-31.8	13.2	-167.4	8.2	-144.7	12.1	30.1	19.2	179.2	4.7
r5-24	*5*	*-89.3*	*10.6*	*-33.2*	*13.5*	*-168.8*	*9.5*	*-139.2*	*9.1*	*31.9*	*21.9*	*-179.4*	*5.9*
r5-25	7	-85.6	18.3	-7.2	14.9	-178.0	4.1	-126.2	17.5	-71.7	13.4	-176.3	5.0
r5-25	*5*	*-79.0*	*11.1*	*-10.8*	*9.5*	*-174.6*	*4.8*	*-131.8*	*12.1*	*-77.9*	*16.7*	*-172.7*	*4.8*
r5-26	6	-101.9	6.5	-7.9	8.4	-179.3	2.6	-132.2	3.2	-53.9	4.0	-173.4	2.2
r5-26	*5*	*-101.4*	*8.0*	*-9.8*	*7.7*	*-178.3*	*1.4*	*-133.8*	*3.0*	*-51.2*	*5.5*	*-173.2*	*2.7*
r5-27	7	71.3	16.0	-140.6	15.7	-178.0	11.5	-90.8	4.7	4.4	12.5	-176.1	5.4
r5-27	*6*	*64.8*	*11.5*	*-133.6*	*11.1*	*175.2*	*6.7*	*-98.8*	*6.8*	*14.0*	*16.5*	*-179.3*	*5.2*
r5-28	6	-59.1	9.9	-32.1	11.9	179.7	3.2	-85.8	10.4	-10.0	16.6	178.5	1.9
r5-28	*6*	*-60.9*	*6.3*	*-25.9*	*6.4*	*178.7*	*2.0*	*-81.0*	*5.2*	*-20.1*	*9.5*	*179.2*	*2.0*
r5-29	5	-105.4	13.2	126.0	14.2	175.4	7.2	62.3	12.7	35.4	19.3	-176.6	1.5
r5-29	*3*	*-113.9*	*10.4*	*118.1*	*19.0*	*-179.7*	*1.4*	*55.5*	*6.5*	*49.8*	*4.2*	*-177.1*	*2.1*
r5-30	5	147.1	11.1	-60.5	4.8	-5.8	5.3	-71.6	9.9	140.2	12.9	175.8	2.7
r5-30	*4*	*152.1*	*4.9*	*-62.7*	*3.3*	*-5.4*	*4.3*	*-70.7*	*8.0*	*144.1*	*9.4*	*176.9*	*3.1*

type	|turns|	ψ_{i+4}		φ_{i+4}		E_{hb} [kcal/mol]	
		∅	±	∅	±	∅	±
r5-23	7	-95.5	16.9	130.4	18.4	-4.3	1.2
r5-23	*5*	*-91.9*	*11.5*	*119.8*	*14.1*	*-4.6*	*0.8*
r5-24	6	-112.1	12.0	136.5	18.7	-4.2	1.3
r5-24	*5*	*-114.6*	*7.8*	*138.7*	*19.9*	*-3.8*	*1.0*
r5-25	7	-170.9	22.9	177.9	14.8	-4.2	0.9
r5-25	*5*	*-161.0*	*20.0*	*171.2*	*14.8*	*-4.1*	*1.7*
r5-26	6	-140.1	5.3	162.6	2.9	-2.2	0.2
r5-26	*5*	*-140.5*	*3.8*	*163.0*	*3.1*	*-2.3*	*0.3*
r5-27	7	-76.3	11.8	147.3	10.2	-3.3	1.1
r5-27	*6*	*-77.0*	*19.3*	*144.2*	*4.4*	*-3.3*	*1.0*
r5-28	6	-143.5	12.4	70.0	10.1	-4.4	1.5
r5-28	*6*	*-145.1*	*12.0*	*66.2*	*6.3*	*-4.6*	*1.0*
r5-29	5	-117.8	17.2	143.9	10.1	-4.2	1.7
r5-29	*3*	*-132.6*	*9.4*	*151.5*	*5.2*	*-4.2*	*0.6*
r5-30	5	-124.1	13.2	144.5	8.3	-4.1	1.1
r5-30	*4*	*-121.1*	*13.1*	*143.9*	*10.0*	*-4.1*	*0.8*

Table A3.18. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 5 turns. Only turn types with at least 50 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*		*position 4*		*position 5*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r5-1												
Ala	476	0.58	55	0.34	236	1.44	81	0.50	20	0.12	84	0.51
Cys	57	0.48	24	1.01	9	0.38	15	0.63	3	0.13	6	0.25
Asp	1,602	2.84	610	5.39	168	1.49	692	6.12	48	0.42	84	0.74
Glu	635	1.01	35	0.28	262	2.07	145	1.15	17	0.13	176	1.39
Phe	171	0.44	49	0.63	27	0.35	26	0.34	17	0.22	52	0.67
Gly	1,781	2.49	120	0.84	59	0.41	49	0.34	1,495	10.45	58	0.41
His	171	0.76	22	0.49	32	0.71	41	0.91	17	0.38	59	1.31
Ile	105	0.19	2	0.02	42	0.38	5	0.05	0	0.00	56	0.51
Lys	503	0.96	37	0.35	165	1.57	87	0.83	31	0.30	183	1.74
Leu	312	0.35	111	0.63	61	0.35	37	0.21	29	0.16	74	0.42
Met	69	0.42	22	0.67	10	0.31	13	0.40	11	0.34	13	0.40
Asn	931	2.35	267	3.37	80	1.01	291	3.67	108	1.36	185	2.33
Pro	224	0.51	0	0.00	222	2.51	1	0.01	0	0.00	1	0.01
Gln	328	0.94	32	0.46	62	0.89	87	1.25	22	0.32	125	1.79
Arg	421	0.87	56	0.58	95	0.98	54	0.56	31	0.32	185	1.90
Ser	703	1.27	155	1.40	208	1.88	142	1.28	35	0.32	163	1.47
Thr	590	1.15	249	2.43	66	0.64	67	0.65	1	0.01	207	2.02
Val	166	0.24	4	0.03	46	0.34	8	0.06	1	0.01	107	0.79
Trp	68	0.51	14	0.52	9	0.33	22	0.82	4	0.15	19	0.71
Tyr	179	0.54	33	0.50	36	0.54	39	0.59	12	0.18	59	0.89
Other	18	0.38	5	0.53	7	0.74	0	0.00	0	0.00	6	0.64
type: r5-2												
Ala	41	0.56	7	0.48	11	0.75	12	0.82	6	0.41	5	0.34
Cys	4	0.38	1	0.47	0	0.00	1	0.47	1	0.47	1	0.47
Asp	96	1.90	4	0.40	34	3.36	32	3.17	24	2.37	2	0.20
Glu	67	1.19	14	1.24	3	0.27	20	1.77	12	1.06	18	1.59
Phe	13	0.38	4	0.58	1	0.14	0	0.00	3	0.43	5	0.72
Gly	18	0.28	0	0.00	12	0.94	4	0.31	2	0.16	0	0.00
His	33	1.63	0	0.00	3	0.74	4	0.99	13	3.22	13	3.22
Ile	51	1.05	40	4.10	0	0.00	1	0.10	4	0.41	6	0.62
Lys	45	0.96	9	0.96	2	0.21	17	1.81	6	0.64	11	1.17
Leu	53	0.67	34	2.16	0	0.00	5	0.32	0	0.00	14	0.89
Met	5	0.34	1	0.34	0	0.00	1	0.34	1	0.34	2	0.68
Asn	72	2.03	0	0.00	30	4.24	12	1.69	25	3.53	5	0.71
Pro	27	0.68	0	0.00	0	0.00	27	3.41	0	0.00	0	0.00
Gln	24	0.77	3	0.48	3	0.48	4	0.64	7	1.12	7	1.12
Arg	31	0.71	3	0.35	1	0.12	6	0.69	14	1.61	7	0.81
Ser	99	2.00	5	0.50	49	4.95	15	1.51	18	1.82	12	1.21
Thr	89	1.94	11	1.20	21	2.29	7	0.76	17	1.85	33	3.60
Val	51	0.84	28	2.30	0	0.00	0	0.00	7	0.58	16	1.32
Trp	19	1.58	4	1.66	0	0.00	1	0.42	3	1.25	11	4.57
Tyr	11	0.37	2	0.34	0	0.00	1	0.17	6	1.01	2	0.34
Other	1	0.24	0	0.00	0	0.00	0	0.00	1	1.19	0	0.00

Table A3.18 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 5 turns. Only turn types with at least 50 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r5-3												
Ala	33	0.56	6	0.51	13	1.10	5	0.42	2	0.17	7	0.59
Cys	2	0.23	1	0.58	0	0.00	0	0.00	0	0.00	1	0.58
Asp	47	1.15	4	0.49	16	1.95	14	1.71	10	1.22	3	0.37
Glu	33	0.72	2	0.22	15	1.64	6	0.65	4	0.44	6	0.65
Phe	16	0.57	4	0.71	2	0.36	2	0.36	5	0.89	3	0.54
Gly	188	3.62	45	4.34	10	0.96	50	4.82	73	7.03	10	0.96
His	15	0.92	4	1.22	3	0.92	5	1.53	1	0.31	2	0.61
Ile	16	0.40	7	0.88	2	0.25	0	0.00	0	0.00	7	0.88
Lys	30	0.79	3	0.39	8	1.05	7	0.92	1	0.13	11	1.45
Leu	48	0.75	30	2.35	3	0.23	2	0.16	1	0.08	12	0.94
Met	10	0.84	2	0.84	3	1.26	1	0.42	1	0.42	3	1.26
Asn	58	2.02	1	0.17	4	0.70	28	4.87	22	3.83	3	0.52
Pro	30	0.93	0	0.00	30	4.67	0	0.00	0	0.00	0	0.00
Gln	27	1.07	4	0.79	0	0.00	7	1.38	3	0.59	13	2.57
Arg	32	0.91	5	0.71	3	0.43	2	0.28	4	0.57	18	2.55
Ser	32	0.80	2	0.25	13	1.62	5	0.62	4	0.50	8	0.99
Thr	24	0.64	8	1.07	4	0.54	1	0.13	1	0.13	10	1.34
Val	24	0.49	4	0.41	5	0.51	0	0.00	0	0.00	15	1.52
Trp	4	0.41	3	1.54	0	0.00	0	0.00	1	0.51	0	0.00
Tyr	21	0.87	3	0.62	4	0.83	3	0.62	5	1.03	6	1.24
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
type: r5-4												
Ala	10	0.28	0	0.00	5	0.71	3	0.43	2	0.28	0	0.00
Cys	4	0.78	4	3.90	0	0.00	0	0.00	0	0.00	0	0.00
Asp	43	1.76	24	4.92	3	0.62	11	2.26	4	0.82	1	0.21
Glu	23	0.84	1	0.18	7	1.28	11	2.02	3	0.55	1	0.18
Phe	11	0.66	1	0.30	0	0.00	1	0.30	9	2.70	0	0.00
Gly	84	2.72	3	0.49	0	0.00	1	0.16	4	0.65	76	12.32
His	20	2.05	8	4.11	3	1.54	1	0.51	8	4.11	0	0.00
Ile	4	0.17	0	0.00	1	0.21	2	0.43	1	0.21	0	0.00
Lys	19	0.84	0	0.00	5	1.11	7	1.55	7	1.55	0	0.00
Leu	13	0.34	0	0.00	3	0.40	4	0.53	6	0.79	0	0.00
Met	7	0.99	2	1.42	2	1.42	0	0.00	2	1.42	1	0.71
Asn	18	1.05	9	2.63	4	1.17	3	0.88	1	0.29	1	0.29
Pro	23	1.20	0	0.00	22	5.76	1	0.26	0	0.00	0	0.00
Gln	9	0.60	0	0.00	2	0.66	3	1.00	4	1.33	0	0.00
Arg	26	1.24	2	0.48	8	1.91	10	2.39	6	1.43	0	0.00
Ser	28	1.17	11	2.30	4	0.84	3	0.63	9	1.88	1	0.21
Thr	40	1.81	17	3.84	5	1.13	11	2.49	7	1.58	0	0.00
Val	14	0.48	0	0.00	6	1.02	6	1.02	2	0.34	0	0.00
Trp	3	0.52	0	0.00	0	0.00	1	0.86	2	1.72	0	0.00
Tyr	11	0.77	0	0.00	2	0.70	3	1.04	5	1.74	1	0.35
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

Table A3.18 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 5 turns. Only turn types with at least 50 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r5-5												
Ala	11	0.37	4	0.67	2	0.34	0	0.00	3	0.51	2	0.34
Cys	3	0.69	1	1.16	0	0.00	0	0.00	0	0.00	2	2.31
Asp	14	0.68	0	0.00	3	0.73	6	1.46	4	0.98	1	0.24
Glu	16	0.70	5	1.09	8	1.75	1	0.22	2	0.44	0	0.00
Phe	10	0.71	2	0.71	1	0.36	0	0.00	4	1.43	3	1.07
Gly	80	3.08	1	0.19	1	0.19	44	8.48	0	0.00	34	6.55
His	8	0.98	1	0.61	3	1.83	0	0.00	4	2.44	0	0.00
Ile	24	1.21	12	3.03	0	0.00	0	0.00	9	2.27	3	0.76
Lys	32	1.68	4	1.05	11	2.89	2	0.53	12	3.15	3	0.79
Leu	15	0.47	3	0.47	2	0.31	1	0.16	3	0.47	6	0.94
Met	9	1.52	1	0.84	0	0.00	1	0.84	3	2.53	4	3.37
Asn	12	0.83	1	0.35	0	0.00	10	3.48	0	0.00	1	0.35
Pro	24	1.49	0	0.00	24	7.47	0	0.00	0	0.00	0	0.00
Gln	14	1.11	5	1.97	3	1.18	0	0.00	6	2.37	0	0.00
Arg	19	1.08	4	1.13	7	1.98	1	0.28	5	1.42	2	0.57
Ser	5	0.25	2	0.50	0	0.00	3	0.75	0	0.00	0	0.00
Thr	9	0.48	0	0.00	2	0.54	0	0.00	4	1.07	3	0.81
Val	31	1.26	21	4.26	1	0.20	0	0.00	9	1.83	0	0.00
Trp	2	0.41	1	1.02	0	0.00	0	0.00	1	1.02	0	0.00
Tyr	7	0.58	1	0.41	1	0.41	0	0.00	0	0.00	5	2.07
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
type: r5-6												
Ala	16	0.56	0	0.00	10	1.74	4	0.69	1	0.17	1	0.17
Cys	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Asp	21	1.05	2	0.50	4	1.00	6	1.51	6	1.51	3	0.75
Glu	22	0.99	4	0.90	7	1.57	6	1.35	0	0.00	5	1.12
Phe	6	0.44	0	0.00	0	0.00	3	1.10	0	0.00	3	1.10
Gly	59	2.34	0	0.00	15	2.98	0	0.00	43	8.53	1	0.20
His	4	0.50	2	1.26	0	0.00	1	0.63	0	0.00	1	0.63
Ile	21	1.09	15	3.90	0	0.00	2	0.52	0	0.00	4	1.04
Lys	28	1.52	2	0.54	6	1.62	6	1.62	3	0.81	11	2.98
Leu	27	0.87	15	2.42	3	0.48	2	0.32	1	0.16	6	0.97
Met	3	0.52	0	0.00	0	0.00	0	0.00	0	0.00	3	2.61
Asn	14	1.00	1	0.36	1	0.36	1	0.36	8	2.87	3	1.07
Pro	26	1.67	0	0.00	0	0.00	26	8.33	0	0.00	0	0.00
Gln	11	0.89	2	0.81	0	0.00	6	2.44	1	0.41	2	0.81
Arg	12	0.70	1	0.29	4	1.17	2	0.58	1	0.29	4	1.17
Ser	17	0.87	3	0.77	11	2.82	0	0.00	1	0.26	2	0.51
Thr	14	0.77	8	2.21	1	0.28	1	0.28	0	0.00	4	1.11
Val	30	1.25	11	2.30	5	1.04	1	0.21	0	0.00	13	2.71
Trp	1	0.21	0	0.00	0	0.00	0	0.00	0	0.00	1	1.05
Tyr	3	0.26	1	0.43	0	0.00	0	0.00	2	0.85	0	0.00
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

Table A3.18 *(continued)*. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 5 turns. Only turn types with at least 50 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		***position 1***		***position 2***		***position 3***		***position 4***		***position 5***	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r5-7												
Ala	12	0.46	2	0.38	4	0.76	0	0.00	3	0.57	3	0.57
Cys	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Asp	19	1.05	0	0.00	7	1.93	5	1.38	5	1.38	2	0.55
Glu	18	0.89	1	0.25	7	1.73	5	1.23	4	0.99	1	0.25
Phe	19	1.53	7	2.83	3	1.21	0	0.00	1	0.40	8	3.23
Gly	37	1.61	0	0.00	3	0.65	32	6.98	1	0.22	1	0.22
His	9	1.24	4	2.76	2	1.38	0	0.00	1	0.69	2	1.38
Ile	17	0.97	10	2.86	0	0.00	0	0.00	0	0.00	7	2.00
Lys	20	1.19	3	0.89	2	0.59	2	0.59	10	2.97	3	0.89
Leu	24	0.85	13	2.30	1	0.18	2	0.35	3	0.53	5	0.89
Met	5	0.95	0	0.00	3	2.86	1	0.95	1	0.95	0	0.00
Asn	31	2.44	0	0.00	20	7.87	9	3.54	1	0.39	1	0.39
Pro	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Gln	15	1.34	2	0.89	1	0.45	3	1.34	5	2.23	4	1.79
Arg	18	1.15	2	0.64	4	1.28	1	0.32	10	3.21	1	0.32
Ser	9	0.51	0	0.00	2	0.56	0	0.00	3	0.84	4	1.13
Thr	8	0.49	0	0.00	1	0.30	0	0.00	6	1.82	1	0.30
Val	24	1.10	10	2.29	0	0.00	0	0.00	6	1.38	8	1.84
Trp	2	0.46	1	1.16	1	1.16	0	0.00	0	0.00	0	0.00
Tyr	18	1.68	6	2.81	0	0.00	1	0.47	1	0.47	10	4.68
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
type: r5-8												
Ala	12	0.50	3	0.62	0	0.00	1	0.21	3	0.62	5	1.04
Cys	10	2.85	2	2.85	0	0.00	5	7.13	0	0.00	3	4.28
Asp	28	1.68	4	1.20	3	0.90	10	3.00	10	3.00	1	0.30
Glu	20	1.08	4	1.08	1	0.27	6	1.61	5	1.34	4	1.08
Phe	8	0.70	3	1.32	0	0.00	0	0.00	0	0.00	5	2.20
Gly	52	2.47	0	0.00	45	10.69	0	0.00	2	0.47	5	1.19
His	4	0.60	2	1.50	0	0.00	0	0.00	2	1.50	0	0.00
Ile	14	0.87	8	2.49	0	0.00	3	0.93	0	0.00	3	0.93
Lys	21	1.36	2	0.65	0	0.00	5	1.62	6	1.94	8	2.59
Leu	17	0.66	8	1.54	1	0.19	3	0.58	0	0.00	5	0.97
Met	3	0.62	1	1.04	0	0.00	0	0.00	2	2.08	0	0.00
Asn	23	1.97	1	0.43	0	0.00	12	5.14	7	3.00	3	1.29
Pro	2	0.15	0	0.00	0	0.00	2	0.77	0	0.00	0	0.00
Gln	7	0.68	0	0.00	1	0.49	1	0.49	2	0.97	3	1.46
Arg	8	0.56	3	1.05	1	0.35	0	0.00	1	0.35	3	1.05
Ser	15	0.92	0	0.00	2	0.61	5	1.53	8	2.45	0	0.00
Thr	8	0.53	1	0.33	1	0.33	3	0.99	1	0.33	2	0.66
Val	11	0.55	7	1.75	0	0.00	0	0.00	1	0.25	3	0.75
Trp	1	0.25	0	0.00	0	0.00	0	0.00	1	1.26	0	0.00
Tyr	16	1.63	7	3.57	1	0.51	0	0.00	5	2.55	3	1.53
Other	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00

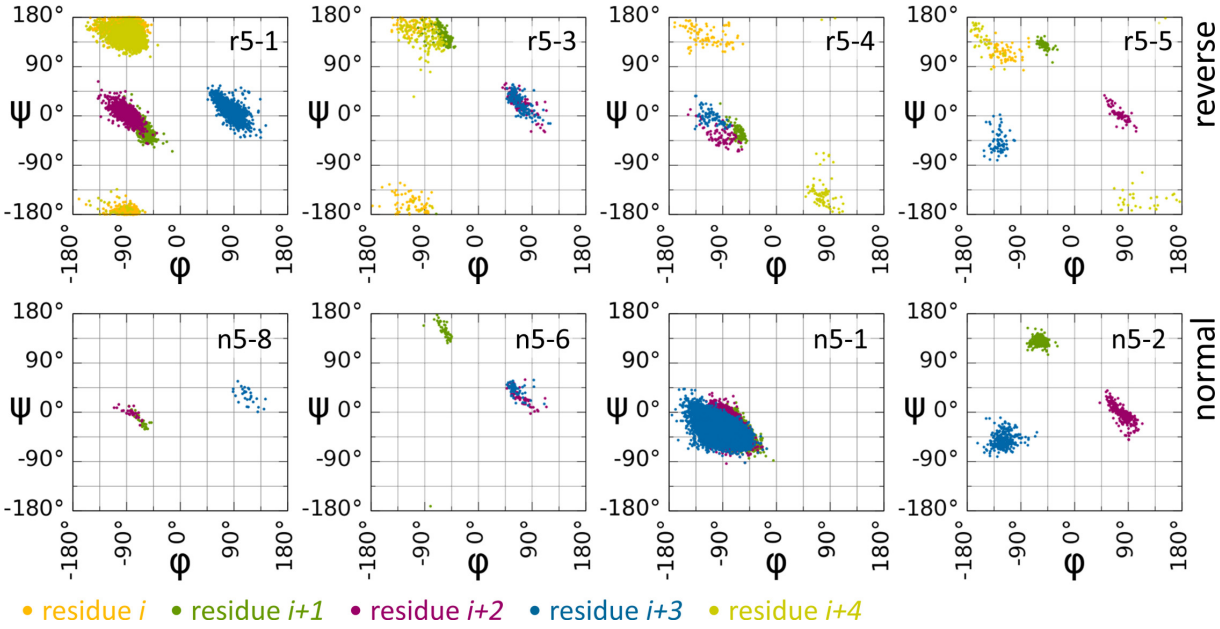


Figure A3.1. Ramachandran plots of the highest occupied related normal 5 and reverse 5 turn types. The relationship between the reverse 5 and normal 5 turn types can be analysed based on the dihedral angle distributions (correspondences between ψ_{i+1}, φ_{i+1}, ψ_{i+2}, φ_{i+2}, ψ_{i+3}, and φ_{i+3}).

REVERSE 6 TURNS

Table A3.19. Frequencies and conformational parameters of residues within different types and at specific positions of reverse 6 turns. Only turn types with at least 50 representatives are covered by this table. The colouring from light to dark green and orange represents the over- and underrepresentation of the residues at significance levels of 5%, 1%, and 0.1%, respectively.

	turn		*position 1*		*position 2*		*position 3*		*position 4*		*position 5*		*position 6*	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***	***n***	***P***
type: r6-1														
Ala	355	0.60	18	0.18	104	1.06	110	1.12	39	0.40	21	0.21	63	0.64
Cys	213	2.49	107	7.49	3	0.21	5	0.35	79	5.53	6	0.42	13	0.91
Asp	765	1.88	438	6.46	62	0.91	92	1.36	72	1.06	61	0.90	40	0.59
Glu	510	1.12	42	0.55	67	0.88	186	2.45	45	0.59	43	0.57	127	1.68
Phe	120	0.43	8	0.17	39	0.84	19	0.41	14	0.30	18	0.39	22	0.47
Gly	587	1.14	13	0.15	25	0.29	39	0.45	18	0.21	465	5.42	27	0.31
His	178	1.10	31	1.14	28	1.03	36	1.33	29	1.07	33	1.22	21	0.77
Ile	171	0.44	19	0.29	44	0.67	35	0.53	15	0.23	2	0.03	56	0.86
Lys	536	1.42	10	0.16	63	1.00	125	1.99	73	1.16	113	1.80	152	2.42
Leu	321	0.51	26	0.25	88	0.83	39	0.37	58	0.55	36	0.34	74	0.70
Met	71	0.60	5	0.25	17	0.87	8	0.41	6	0.31	10	0.51	25	1.27
Asn	545	1.91	211	4.44	33	0.69	53	1.11	100	2.10	121	2.55	27	0.57
Pro	258	0.81	0	0.00	241	4.54	16	0.30	1	0.02	0	0.00	0	0.00
Gln	266	1.06	15	0.36	28	0.67	53	1.27	31	0.74	68	1.62	71	1.69
Arg	408	1.17	26	0.45	71	1.22	63	1.08	60	1.03	83	1.42	105	1.80
Ser	520	1.30	104	1.56	53	0.80	97	1.46	153	2.30	38	0.57	75	1.13
Thr	600	1.63	14	0.23	69	1.12	69	1.12	318	5.17	2	0.03	128	2.08
Val	241	0.49	29	0.36	72	0.88	57	0.70	9	0.11	5	0.06	69	0.85
Trp	47	0.49	7	0.43	8	0.50	14	0.87	7	0.43	0	0.00	11	0.68
Tyr	117	0.49	16	0.40	20	0.50	24	0.60	13	0.33	13	0.33	31	0.78
Other	17	0.50	2	0.35	6	1.06	1	0.18	1	0.18	3	0.53	4	0.71
type: r6-2														
Ala	43	0.50	9	0.62	13	0.90	9	0.62	1	0.07	5	0.35	6	0.42
Cys	7	0.56	1	0.48	1	0.48	1	0.48	2	0.95	1	0.48	1	0.48
Asp	163	2.72	26	2.60	16	1.60	102	10.21	3	0.30	11	1.10	5	0.50
Glu	45	0.67	2	0.18	13	1.16	7	0.63	5	0.45	9	0.81	9	0.81
Phe	22	0.54	2	0.29	2	0.29	0	0.00	5	0.73	0	0.00	13	1.91
Gly	130	1.72	8	0.63	8	0.63	0	0.00	103	8.15	3	0.24	8	0.63
His	15	0.63	1	0.25	0	0.00	1	0.25	3	0.75	2	0.50	8	2.00
Ile	8	0.14	2	0.21	0	0.00	0	0.00	0	0.00	1	0.10	5	0.52
Lys	72	1.30	2	0.22	10	1.08	2	0.22	10	1.08	27	2.91	21	2.27
Leu	45	0.48	5	0.32	9	0.58	7	0.45	7	0.45	5	0.32	12	0.77
Met	8	0.46	2	0.69	1	0.35	0	0.00	1	0.35	2	0.69	2	0.69
Asn	69	1.64	24	3.43	8	1.14	13	1.86	6	0.86	14	2.00	4	0.57
Pro	44	0.94	0	0.00	44	5.62	0	0.00	0	0.00	0	0.00	0	0.00
Gln	25	0.68	0	0.00	4	0.65	2	0.32	6	0.97	6	0.97	7	1.13
Arg	56	1.09	2	0.23	10	1.16	3	0.35	6	0.70	21	2.45	14	1.63
Ser	143	2.44	65	6.64	20	2.04	9	0.92	6	0.61	31	3.17	12	1.23
Thr	61	1.12	15	1.66	3	0.33	8	0.88	0	0.00	18	1.99	17	1.88
Val	23	0.32	0	0.00	2	0.17	0	0.00	0	0.00	10	0.83	11	0.92
Trp	10	0.70	0	0.00	3	1.26	2	0.84	0	0.00	0	0.00	5	2.10
Tyr	17	0.48	2	0.34	1	0.17	1	0.17	4	0.68	2	0.34	7	1.19
Other	2	0.40	0	0.00	0	0.00	1	1.20	0	0.00	0	0.00	1	1.20

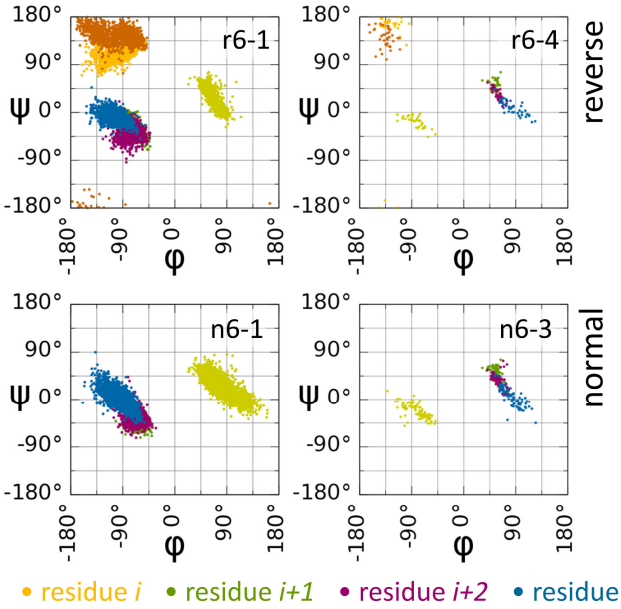


Figure A3.2. Ramachandran plots of the highest occupied related normal 6 and reverse 6 turn types. The relationship between the reverse 6 and normal 6 turn types can be analysed based on the dihedral angle distributions (correspondences between ψ_{i+1}, φ_{i+1}, ψ_{i+2}, φ_{i+2}, ψ_{i+3}, φ_{i+3}, ψ_{i+4}, and φ_{i+4}).

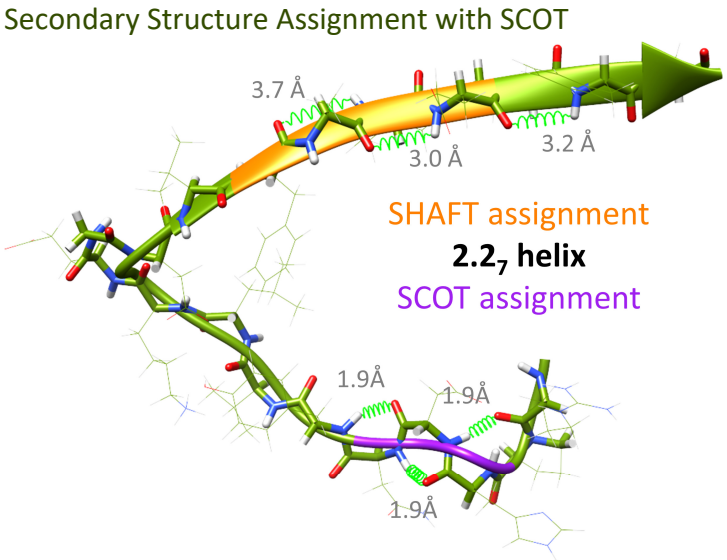


Figure A3.3. 2.2$_7$-helix assignments by SCOT and SHAFT. Residues 171 to 190 of methionine aminopeptidase from *Pyrococcus furiosus* are presented in green. The cartoon representation shows the presence of a β-strand from residue 183 to 190. The SHAFT- (orange) and SCOT- (purple) assigned left-handed 2.2$_7$-helices are highlighted together with the hydrogen bonds (green springs) of the n3-1 turns which were used for the classification. The acceptor hydrogen distances are given next to the hydrogen bonds.

Table A3.20. The assignment of β-strands and left-handed helices by different SSAMs. The consensus between ASSP, DISICL, MKDSSP, SCOT, and SEGNO is given based on a per-residue assignment comparison. Some methods do not support the classification of the SSE (n/a). The highest consensus with SCOT is highlighted in bold.

β-strands	ASSP	DISICL	MKDSSP	SCOT	SEGNO
ASSP	1	0.5001	0.4859	0.4788	0.5439
DISICL		1	0.535	0.5158	0.6565
MKDSSP			1	**0.8982**	0.6065
SCOT				1	0.5916
SEGNO					1
PPII helices	ASSP	DISICL	MKDSSP	SCOT	SEGNO
ASSP	1	0.0936	n/a	0.1644	0.2021
DISICL		1	n/a	0.0735	0.1708
MKDSSP			n/a	n/a	n/a
SCOT				1	**0.1687**
SEGNO					1
left-handed helices	ASSP	DISICL	MKDSSP	SCOT	SEGNO
ASSP	1	0.6704	0.3004	0.6501	n/a
DISICL		1	0.3761	**0.8534**	n/a
MKDSSP			1	0.3730	n/a
SCOT				1	n/a
SEGNO					n/a

Table A3.21. Geometric characteristics of SCOT-assigned PPII helices. The angular parameters twist, vtor (virtual torsion angle), and the bending angle (BDA) as well as the dihedral angles are given in degrees. The radius (r) and the rise are given in Å. Mean values of all parameters and the respective standard deviations in brackets are given. The given B-factor was scaled to ensure comparability between multiple protein structures.

	twist	**rise**	**r**	**vtor**	**BDA**	***φ***	***ψ***	***ω***	**B-factor**	**Length**
SCOT	9.0	-2.8	1.4	-111.0	41.8	-78.9	143.0	178.0	0.3	4.2
	(0.5)	(1.0)	(0.3)	(29.3)	(35.8)	(22.3)	(23.4)	(5.1)	(1.0)	(1.5)
ASSP	9.0	-3.0	1.4	-110.7	13.7	-77.2	141.7	177.7	0.2	3.3
	(0.4)	(0.1)	(0.2)	(16.9)	(7.7)	(21.7)	(18.4)	(5.7)	(1.0)	(0.7)
DISICL	8.9	-2.4	1.4	-111.5	20.4	-83.3	137.2	178.3	0.1	2.3
	(0.5)	(1.6)	(0.3)	(34.2)	(9.6)	(34.6)	(49.0)	(11.2)	(1.0)	(0.6)
SEGNO	9.1	-2.9	1.3	-112.8	29.3	-77.4	144.8	177.9	0.3	3.6
	(0.6)	(0.6)	(0.3)	(25.1)	(21.4)	(21.0)	(16.3)	(10.4)	(1.0)	(1.0)

Table A3.22. The results of different SSAMs for the left-handed helices in the dataset of Novotny and Kleywegt[1]. The results for DSSP are only given for at least two residue long helices. DISICL assigns left-handed helices, but no distinct types. The helices are grouped according to their type: α, 3_{10}, differently assigned by the algorithms (≠), and rarely assigned left-handed helices (?). Some helices were not found by the methods (n/a).

	PDB-ID. Chain	Manual[1]	DSSP[2]	ASSP[3]	DISICL[4]	SCOT	CATH domain
3_{10}	1aut.L	101-104	101-103	101-104	101-104	101-104	2.10.25.10
	1b9w.A	52-55	52-54	n/a	52-55	52-55	2.10.25.10
	1b9w.A	n/a	83-84	n/a	83-85	n/a	2.10.25.10
	1g2l.B	258-261	258-260	259-261	258-261	258-261	2.10.25.10
	1h21.A	n/a	39-40	n/a	n/a	n/a	–
	1jv1.A	182-185	182-184	182-185	182-185	182-185	3.90.550.10
	1kdg.A	n/a	450-451	n/a	n/a	n/a	3.30.410.10
	1kdg.A	532-535	532-534	532-535	532-535	532-535	3.30.410.10
	1kli.L	94-97	94-96	95-97	94-97	94-97	2.10.25.10
	1n1i.A	57-60	57-59	57-60	57-60	57-60	2.10.25.10
	1n1i.A	n/a	88-89	n/a	88-90	88-90	2.10.25.10
	1ob1.C	52-55	52-54	53-55	52-55	52-55	2.10.25.10
	1ob1.C	n/a	87-88	n/a	87-89	87-89	2.10.25.10
	1pb5.A	28-32	29-31	29-32	28-32	29-32	–
	1rfn.B	91-94	90-93	92-94	91-94	91-94	2.10.25.10
	2gsa.A	65-68	65-67	65-67	65-67	65-67	3.90.1150.10
	2gsa.A	n/a	24-25	n/a	n/a	n/a	3.90.1150.10
≠	1h21.A	77-81 (3_{10})	77-78 (3_{10})	77-81 (π)	n/a	n/a	–
	1hxx.A	143-146 (3_{10})	143-145 (α)	143-146 (α)	143-146	n/a	2.40.160.10
	1j9q.A	105-108 (3_{10})	105-107 (3_{10})	105-108 (α)	105-108	105-107 (3_{10})	2.60.40.240
	1nif.A	105-108 (3_{10})	105-107 (3_{10})	105-108 (α)	105-108	n/a	2.60.40.240
	1oe1.A	99-102 (3_{10})	99-101 (3_{10})	99-102 (α)	99-102	n/a	2.60.40.240
	1qj5.A	50-53 (3_{10})	n/a	50-53 (α)	n/a	n/a	3.90.1150.10
	2oat.A	83-86 (3_{10})	83-85 (3_{10})	83-86 (α)	83-86	83-86 (3_{10})	3.90.1150.10
α	1ak0.A	131-134	130-133	131-134	131-134	n/a	1.10.575.10
	1bd0.A	40-44	40-43	40-44	40-44	40-44	3.20.20.10
	1bnl.A	135-138	135-137	135-138	135-138	n/a	3.10.100.10
	1bqb.A	223-226	223-225	223-226	223-226	223-226	1.10.390.10
	1dy2.A	207-210	207-209	207-210	207-210	207-210	3.10.100.10
	1hzm.A	61-64	61-63	60-63	61-64	n/a	3.40.250.10
	1kdg.A	n/a	552-553	n/a	n/a	n/a	3.30.410.10
	1koe.A	266-269	266-268	266-269	266-269	266-269	3.10.100.10
	1kws.A	298-301	298-300	298-301	298-301	298-301	3.90.550.10
	1npc.A	227-230	227-229	227-230	227-230	227-230	1.10.390.10
	1ohv.A	70-73	70-72	70-73	70-73	70-73	3.90.1150.10
	8tln.E	226-229	226-228	226-229	226-229	226-229	1.10.390.10
?	1mzr.A	191-194 (3_{10})	n/a	n/a	n/a	n/a	3.20.20.10
	1ptm.A	211-216 (3_{10})	211-215 (3_{10})	n/a	211-213	211-215 (3_{10})	3.40.718.10

REFERENCES

[1] Novotny, M.; Kleywegt, G. J. A survey of left-handed helices in protein structures. *J. Mol. Biol.* **2005**, *347* (2), 231–241.

[2] Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22* (12), 2577–2637.

[3] Kumar, P.; Bansal, M. Identification of local variations within secondary structures of proteins. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2015**, *71* (Pt 5), 1077–1086.

[4] Nagy, G.; Oostenbrink, C. Dihedral-based segment identification and classification of biopolymers I: Proteins. *J. Chem. Inf. Model.* **2014**, *54* (1), 266–277.

APPENDIX 4: Trypanothione Synthetase

TRYPANOTHIONE SYNTHETASE SEQUENCES

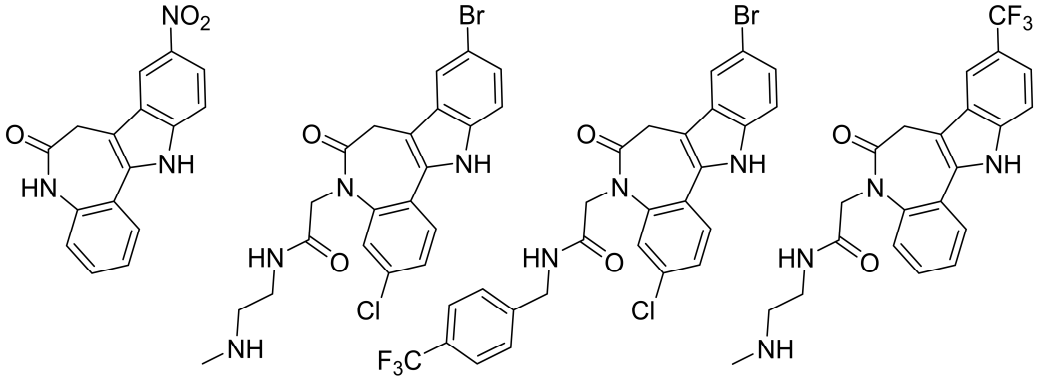
Table A4.1. Sequences of the different TryS enzymes that were used for comparative modelling. The sequences are given with their notation used in the manuscript, their UniProt accession number, and the respective host organism in brackets. The sequence residues that were modelled for structure analyses are underlined.

Protein Name	Sequence
*Cf*TryS	>O60993 (*Crithidia fasciculata* (strain HS6)) MASAERVPVSFNKPGRVPFGEVQGYAPGHIPAYSNKHDHFFSGERSIDDNVFCGFKYQCV EFARRWLLERKGLVLPDVNWACHIFKLKNVKDAATAEEVPVIAVRNGTEAKPEPDTLIIY PSSDVNTVGHVGAITEVGDDYVCIADQNYRFHKWEASYSYKLKLQHKDGVWTIIDDIDPN DVEIPLGWVTFPGYENRPEGAAPPALHPSLHFQPPEEPYLVRKTYEPTETKANWLDLNDP AEKLFVEEFGMDVSRSRLEETTVNYYECDHEFHLRCIAYGTQLHDYFMEATAQVINDDER LRIFKIPEELWPRMRHSWKYQQTYISGRFDFAYNNETHQMKCFEYNADSASTLLECGRIQ QKWAESVGLDKEGTRGSGWAVERNLRTAWATCGATGRVHFLVDDEKEEQYTALYCLQAAE AVGLEGKLCVMYDEFRFNEEGYVVDSDGVRVRNIWKTWMWESAISDYFAAQAERGADWKA TPADKVRLCDLMLGKDWDILYFEPMWKLIPSNKAILPIIYHNHPDHPAILRAEYELTDEL RAVGYARKPIVGRVGRNVTITDGTGEVHAESGGNFGERDMIYQELFSLTKQDGYYAIIGG WMLGDAFSGTGIREDKSIITGLDSPFAAIRIKINAIPRPLTHKDLDKLAEDE
*Li*TryS	>A4I2Z3 (*Leishmania infantum* (strain JPCM5)) MSSLPRASVSFNRPGHIPFGAVQGYAPGGVPAYSNKHDHYFSGERSIEGNIFFGFKYQCV EFARRWLLVRKGLVLPDVNWACHIFQLKEVRDAATAESFAVLPVRNGTGTKPEADALLVY PSTATNPVGHVGAITEVGDDYVCVADQNYRFHKWESSYAYKLKLDHRDGIWTIIDDIDAD EIEIPLGWVTFPGRANRPEGAPPVALHPSLHFNKPPKPYLLRRNFLPTESKANWLDMKNP AERLFVEEFGMDVSRTRLEEKAVSYYESNHEFHLRCIAYGTQLHDIFMEATAQVIESDEK LRLFAIPEEFWPRIRHSWKYQQTYISGRFDFAFNNETGEVKCFEYNADSASTLLECGRIQ QKWAESVGLDKQGTRGSGFAVERNLKMAWANSGATGRVHFCVDEEKEEQYTALYCMQAAE AAGLEGKLCVLFDEFRFDDNGHVVDSDGVRVRNVWKTWMWESAITDYYAAREERGENWKP SPKEKVRLCDLLLGDDWEILYFEPMWKVIPSNKAILPMIYHNHPEHPAILKAEYELTDEL RKHGYAKKPIVGRVGSNVTITSGDGEVHAESGGNYGKRNMIYQQLFELKKQDDYYAIIGG WMIGDAFSGTGIREDKSVITGVDSPFAAIRIKTDKLPHPVTHKDIDEMAEDE
*Lm*TryS	>Q711P7 (*Leishmania major* (strain Friedlin)) MSSLQRASVSFNKPGHIPFGAVQGYAPGGVPAYSNKHDHYFSGERNIEDNIFFGFKYQCV EFARRWLLVRKGLLLPDVNWACHIFQLKEVRDAATTESFAVLQVRNGTTTKPEADALLVY PSTDANPVGHVGTITEVGDDYVCVADQNYRFHKWESSCAYKLKLDHRDGIWTIIDDIDAD EIEIPLGWLTFPGRANRPEGAPPVALHPSLHFKEPPKPYLLRRNFLPTESKANWLDMNNP AERLFVEEFGMDVSRTRLEEKVVSYYESNHEFHLRCVAYGTQLHAIFMEATAQVIESDEK LRLFAIPEEFWPRIRHSWKYQQTYISGRFDFAFNNETGEVKCFEYNADSASTLLECGLIQ QKWAESVGLDKQDTRGSGFAVERNLKMAWANSGATGRVHFCVDEEREEQYTALYCMQAAE AVGLEGKLCILFDEFRFDDNGHVVDSDGVRVRNVWKTWMWESAITDYYAAREERGENWKP SPKDKVRLCDLLLGDDWEILYFEPMWKVIPSNKAILPMIYHNHPEHPAILKAEYELTDEL RKHGYAKKPIVGRVGSNVIITSGDGVVHAESGGKYGKRNMIYQQLFELKKQDDYYAIIGG WMIGDAFSGTGIREDKSVITGVDSPFAAVRIKTDKLPHPVTLKDIDKMAEDE
*Tbb*TryS	>Q86MC2 (*Trypanosoma brucei brucei*) MTKSALADTKEEPHVPFGEIQGYTPCGVPAYSNGHDGFFSGERSIDGNLFCGFKYQCVEF ARRWLYEAKGLVLPDVNWAAHIFDLTEVHDASTATPVPCVKVSNGTAAKPVADSLLIYAV NEDAPWGHVAVITEVGDKWVRIADQNHRFHKWKGTYSAELLLKHEGGVWTVEDHAAEGIF VPLGWVTFPSRPNRNPKEPLVLHESLYFKQPEKPFLRRVVFTPENRKTDWLDLTNEAEAE FYKTFGKEATRGGVYESCYYLMNRELYLNCVRYGTQLHSFFLEATKQVLESDDKLRRFRI PEEYWPRIRHSWKTQPHAITGRFDFVFDENTQEFKCFEYNADSASTLLECAVIQEKWANS VGLDDNATRSSGKFMPQTLVRAWEMTGLKGRVHFLVDDDGEERYTALYVMEKAREAGIDA KLCVMFDEFHFDEKGAVVDSDGIPATAVWKTWMWETAISDHQAAREQRGAEWKPTPKDKV RLCDILLGNNWDVRVFEPMWKLIPSNKAILPIIYNNHPDHPAILPASYELTDELRRTGYA KKPIVGRVGRNVTVTEPDGKVLAESDGNFSNRDMVYQQLFRIPKRGDYYAILGGWMLGDT YSGTGVREDKKLTTGLESPFGPVRIQM

Table A4.1 *(continued)*. Sequences of the different TryS enzymes that were used for comparative modelling. The sequences are given with their notation used in the manuscript, their UniProt accession number, and the respective host organism in brackets. The sequence residues that were modelled for structure analyses are underlined.

Protein Name	Sequence
*Tb*TryS	>Q586P2 (*Trypanosoma brucei brucei* (strain 927/4 GUTat10.1)) MTKSALADTKEEPHVPFGEIQGYTPCGVPAYSNGHDGFFSGERSIDGNLFCGFKYQCVEF ARRWLYEAKGLVLPDVNWAAHIFDLTEVHDASTATPVPCVKVSNGTAAKPVADSLLIYAV NEDAPWGHVAVITEVGDKWVRIADQNHRFHKWKGTYSAELLLKHEGGVWTVEDHAAEGIF VPLGWVTFPSRPNRNPKEPLVLHESLYFKQPEKPFLRRVVFTPENRKTDWLDLTNEAEAE FYKTFGKEATRGGVYESCYYLMNRELYLNCVRYGTQLHSFFLEATKQVLESDDKLRRFRI PEEYWPRIRHSWKTQPHAITGRFDFVFDENTQEFKCFEYNADSASTLLECAVIQEKWANS VGLDDNATRSSGKFMPQTLVRAWEMTGLKGRVHFLVDDDGEERYTALYVMEKAREAGIDA KLCVMFDEFHFDEKGAVVDSDGIPVTAVWKTWMWETAISDHQAAREQRGAEWKPTPKDKV RLCDILLGNNWDIRVFEPMWKLIPSNKAILPIIYNNHPDHPAILPASYELTDELRRTGYA KKPIVGRVGRNVTVTEPDGKVLAESDGNFSNRDMVYQQLFRIPKRGDYYAILGGWMLGDT YSGTGVREDKKLITGLESPFGPVRIQM
*Tc*TryS	>Q9GT49 (*Trypanosoma cruzi* (strain CL Brener)) MPTLQSLAVPFGCVQGYAPGGIPAYSNKHESYFSGERSIDGNLFCGFKYQCVEFARRWLF ERKSLVLPDVDWAVHIFNLKEVSDARTGKPVRCVAIRNGTAAKPVVDSLLIYPSDDYSPV GHVAAITEVGDKWVRIADQNHRFHKWDANYAAELPLIHEKGVWTILDPLEDEVLKPLGWV TFPDTPDRNPNEPLVLHESLHFKRGELPTLRRLTFTPTSREKDWLDLTNEAEAYFADVCG IDVKNPKLEKASYYQMNRELYLDCAKYGNQLHQMFLEATKFVLGSDELLRLFCIPEEYWP RLRHSWETQPHAITGRFDFAFDEDTQQFKCFEYNADSASTLLECGVIQQKWARSVGLDDG TTYSSGSLVSSRLQLAWEMAEVTGRVHFLIDNDDEEHYTALYVMQHASAAGLETKLCVLF DEFHFDENGVVVDSDGVAVTTVWKTWMWETAIADHQKARVQRGNDWRPTPKDEVRLCDIL LGPNWDLRVFEPMWKIIPSNKAILPIIYNKHPDHPALLRASYELTVELQRTGYVRKPIVG RVGRNVTVTEASGDIAAKSDGDFSDRDMVYQELFRLPERDGYYAILGGWVIGDVYCGTGV REDKTIITGLESPFSALRVYQGAQRRPLTHEDLDKAEAAAVGGGLKT

TRYPANOTHIONE SYNTHETASE INHIBITORS



	alsterpaullone	*MOL2008*	*MOL2008_aryl*	*FS-554*
***Li*TryS**	n/a	0.15 ± 0.06 µM[a]	55.4 ± 3.7%[b]	0.35 ± 0.05 µM[b]
***Tc*TryS**	n/a	40.5 ± 5.9%[a]	9.0 ± 2.0 µM[b]	44.5 ± 3.8%[b]
***Tb*TryS**	n/a	59.0 ± 6.0%[a]	63.5 ± 7.2%[b]	70.0 ± 4.9%[b]

Figure A4.1. Structures and TryS activities for differently subsitituted paullones scaffolds. For the alsterpaullone (9-nitro-7,12-dihydrobenzo[2,3]azepino[4,5-*b*]indol-6(5*H*)-one) which was co-crystallized with GSK-3β (PDB-ID 1q3w), IC_{50} values are not published (n/a). Percentages give the percent inhibition at a compound concentration of 30 µM whereas concentrations give the IC_{50} value for the corresponding targets. (a) The biochemical data for compound MOL2008 (2-(9-bromo-3-chloro-6-oxo-7,12-dihydrobenzo[2,3]azepino-[4,5-*b*]indol-5(6*H*)-yl)-*N*-(2-(methylamino)ethyl)acetamide) was taken from Benítez and co-workers[1]. (b) The biochemical data for compound MOL2008_aryl (2-(9-bromo-3-chloro-6-oxo-7,12-dihydrobenzo[2,3]azepino[4,5-*b*]indol-5(6*H*)-yl)-*N*-(4-(trifluoromethyl)benzyl)acetamide) and FS-554 (*N*-(2-(methylamino)ethyl)-2-(6-oxo-9-(trifluoromethyl)-7,12-dihydrobenzo[2,3]azepino[4,5-*b*]indol-5(6*H*)-yl)acetamide) are provided according to the results of Orban *et al.*[2].

MD SIMULATION STUDIES

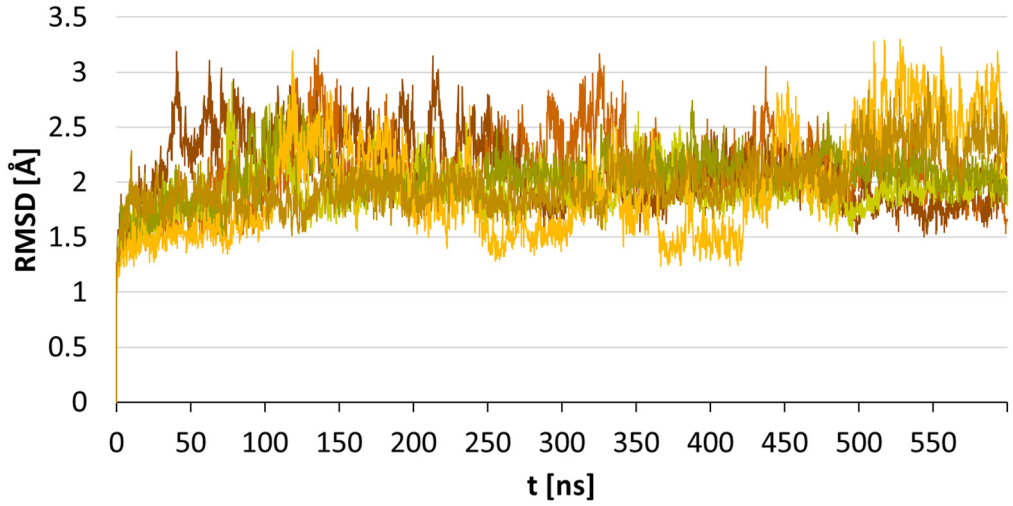


Figure A4.2. RMSD during the MD simulations of *Lm*TryS and *Li*TryS (residues 5 to 652). The RMSD over all Cα atoms of residues 5 to 652 is given in 0.2 ns time steps for the MD simulations of the ligand-free structures of *Lm*TryS (red) and *Li*TryS (dark red), the ATP-bound structures of *Lm*TryS (green) and *Li*TryS (dark green), and the *Lm*TryS and *Li*TryS structures (yellow and dark yellow) with the substrates of the second reaction step (ATP, GSH, and GSP).

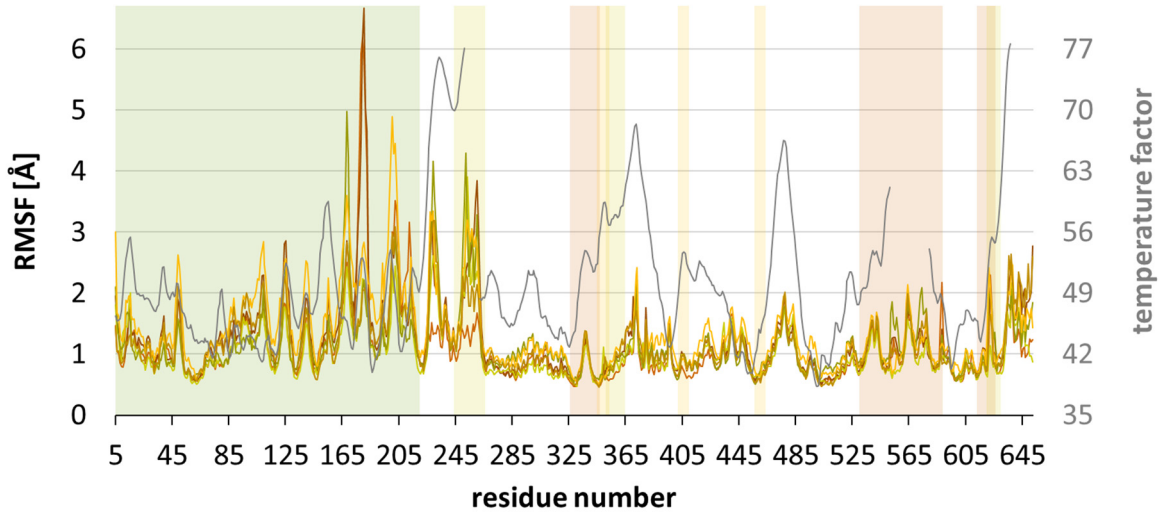


Figure A4.3. RMSF plots for different MD simulations of *Lm*TryS and *Li*TryS. The RMSF of the residues' Cα atoms is given for the 600 ns MD simulations of the ligand-free structures of *Lm*TryS (red) and *Li*TryS (dark red), the ATP-bound structures of *Lm*TryS (green) and *Li*TryS (dark green), and the *Lm*TryS and *Li*TryS structures (yellow and dark yellow) with the substrates of the second reaction step (ATP, GSH, and GSP). The structures were aligned to the first MD snapshot with respect to the Cα atoms of resiudes 5 to 652. For comparison purposes, the B-factor of the residues' Cα atoms in the X-ray structure with the PDB-ID 2vpm (chain A) is given in grey (secondary axis). Residues of the amidase domain, ATP-, GSH-, and GSP-binding residues are highlighted by green, red, yellow, and light green bars, respectively.

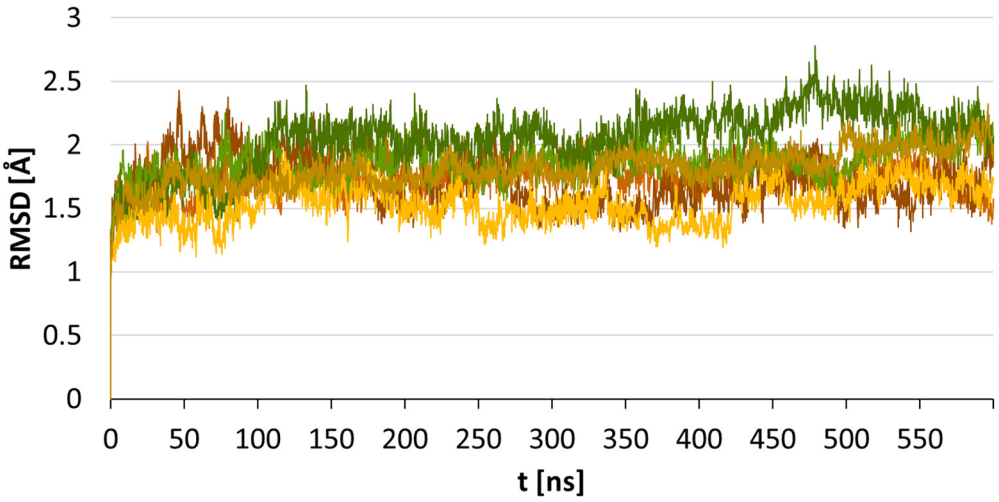


Figure A4.4. RMSD during the MD simulations of *Lm*TryS and *Li*TryS (residues 220 to 652). The RMSD over all Cα atoms of residues 220 to 652 is given in 0.2 ns time steps for the MD simulations of the ligand-free structures of *Lm*TryS (red) and *Li*TryS (dark red), the ATP-bound structures of *Lm*TryS (green) and *Li*TryS (dark green), and the *Lm*TryS and *Li*TryS structures (yellow and dark yellow) with the substrates of the second reaction step (ATP, GSH, and GSP).

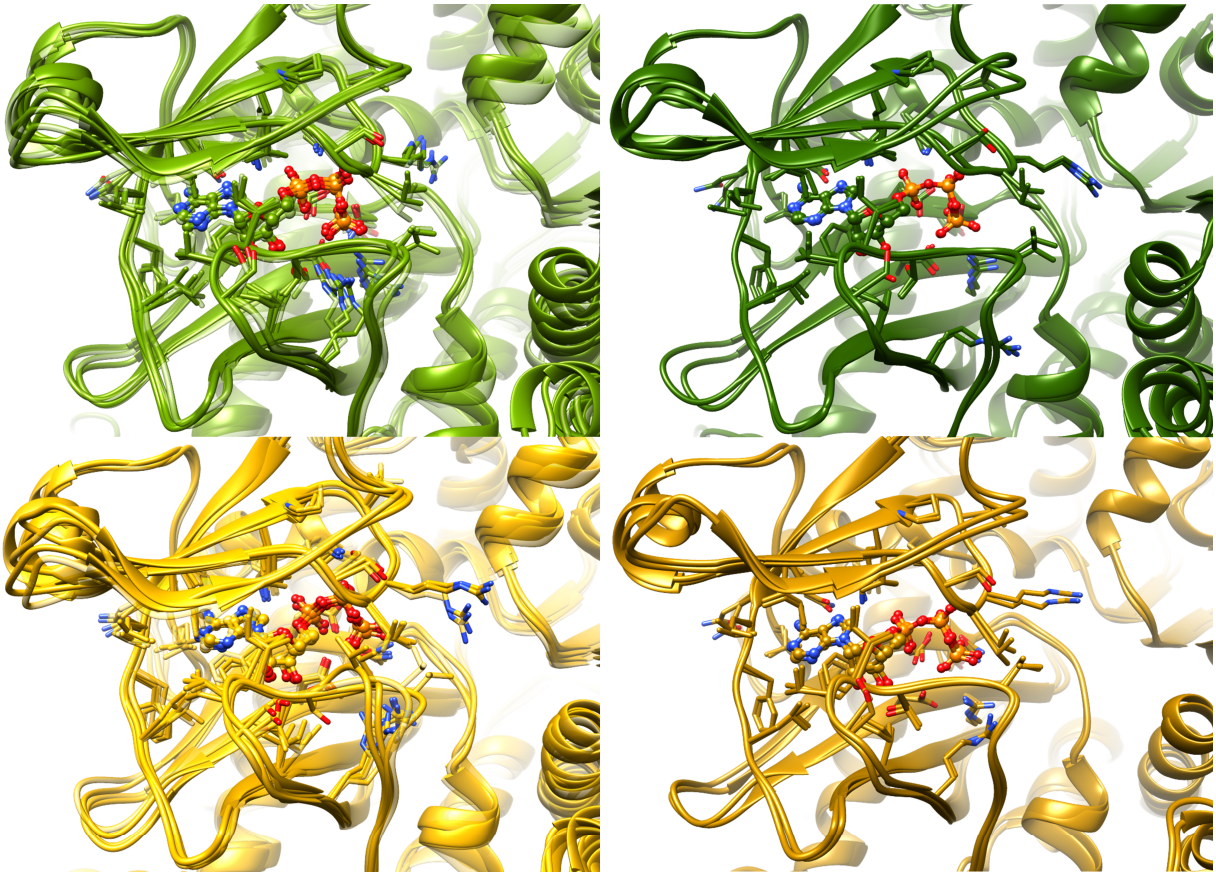


Figure A4.5. Representative structure of *Lm*TryS (left) and *Li*TryS (right) used for docking and binding site comparison studies. The representatives were extracted by applying a hierarchical agglomerative approach with average linkage for an RMSD-based clustering of the snapshots. Clustering was performed focusing on the residues in a 5 Å radius of the TryS ATP binding site (shown in the stick representation). The minimum distance between clusters was set to be greater than 1 Å. The representatives are aligned according to a 8 Å environment of the bound ATP molecule (ball-and-stick representation). The structures on the top are snapshots from the simulations in the presence of ATP. The representatives extracted from the simulation in the presence of ATP, GSH, and GSP (only ATP is shown) are given on the bottom. The representative of the lowest occupied clusters are shown in semi-transparent colours.

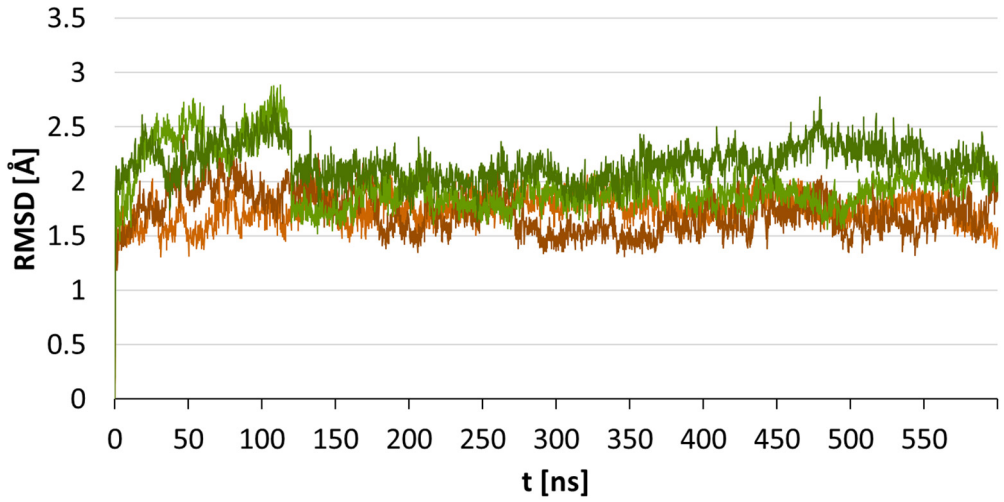
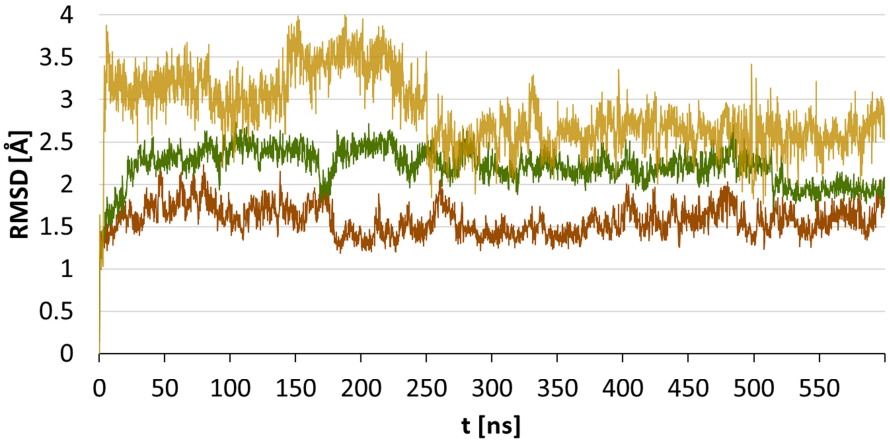


Figure A4.6. RMSD during the MD simulations of *Lm*TryS, *Li*TryS (residues 220 to 652), *Tc*TryS (residues 210 to 638), and *Tb*TryS (residues 216 to 627). The RMSD over all Cα atoms of residues 220 to 652 is given in 0.2 ns time steps for the MD simulations of the ligand-free structures of *Lm*TryS (red), *Li*TryS (dark red), *Tc*TryS (green), and *Tb*TryS (dark green).

Table A4.2. Hydrogen bond analysis results for the MD simulations of *Tc*TryS and *Tb*TryS. The Arg residue which is equivalent to Glu587 in *Lm*TryS and *Li*TryS undergoes several hydrogen bonds during the simulations. Hydrogen bonding interactions which are observed in at least 5% of the 600 ns MD simulation are given. Equivalent residues in both enzymes after sequence alignment are given in the same row. The asterisk denotes mainchain hydrogen bond interactions.

***Tc*TryS (Arg575)**		***Tb*TryS (Arg581)**	
hydrogen bond acceptor	percentage	hydrogen bond acceptor	percentage
Asp322	83.17%	Asp328	93.33%
Thr325	7.17%	Thr331	82.50%
Gln327	5.13%	Glu333	34.83%
Glu550	*217.83%*		
*Thr531**	*13.17%*		
Glu578	*7.67%*		

PREDICTION OF THE PAULLONE BINDING MODE



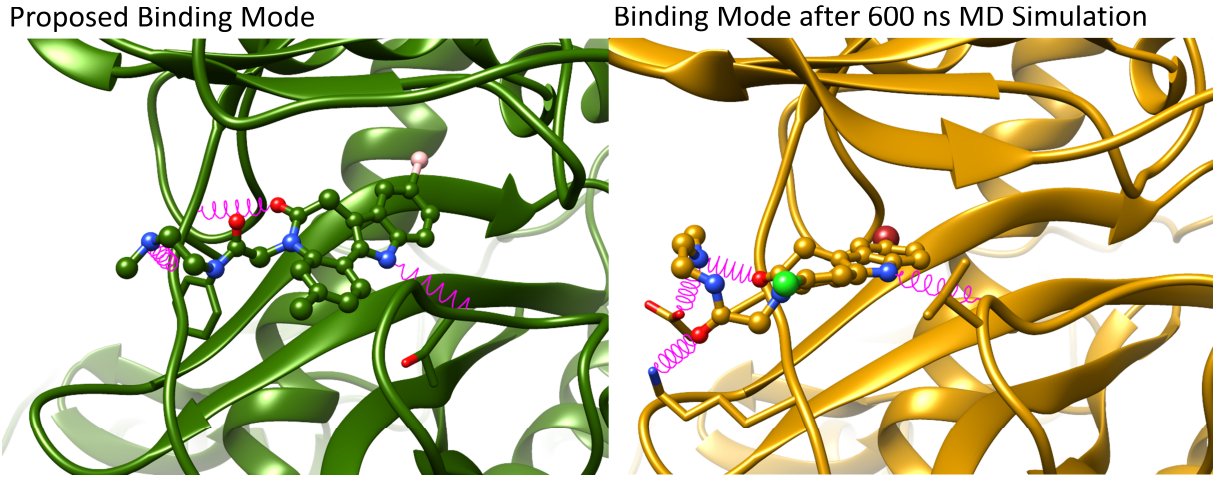


Figure A4.7. RMSD during the MD simulations of *Li*TryS (residues 220 to 633) and the initially proposed binding mode of FS-554. Top: The RMSD over the Cα atoms of residues 220 to 633 (the highly flexible C-terminal residues 633 to 652 were excluded in this analysis) is given in 0.2 ns time steps for the MD simulations of the ligand-free structures of *Li*TryS (dark red) and the *Li*TryS in complex with MOL2008 in the proposed binding mode (dark green). The RMSD of the inhibitor MOL2008 (dark yellow) is also given. Botton: The starting structure of the complex is shown in green whereas the final equilibrated complex is shown in yellow (MOL2008 is given in the ball-and-stick represenation). Hydrogen bonding interactions are indicated by pink springs.

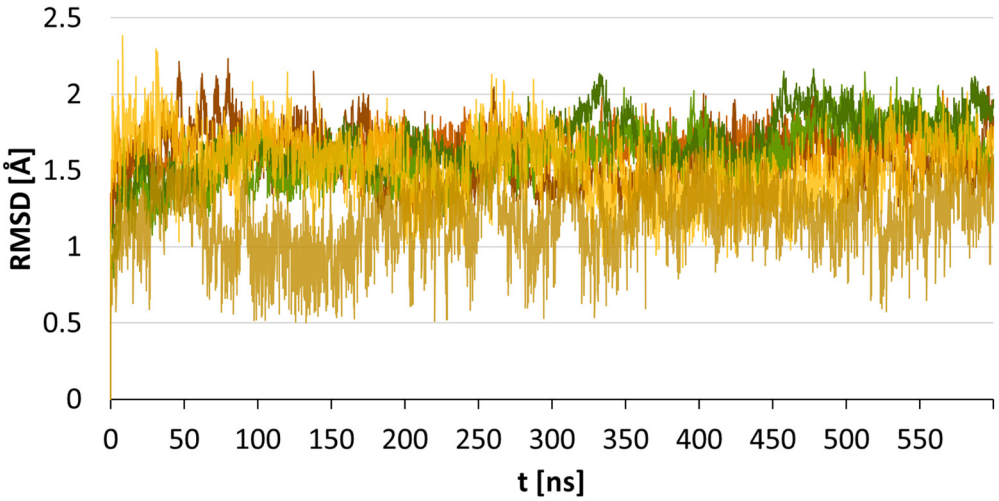


Figure A4.8. RMSD during the MD simulations of *Lm*TryS and *Li*TryS (residues 220 to 633) in complex with FS-554. The RMSD over the Cα atoms of residues 220 to 633 (the highly flexible C-terminal residues 633 to 652 were excluded in this analysis) is given in 0.2 ns time steps for the MD simulations of the ligand-free structures of

*Lm*TryS (red) and *Li*TryS (dark red), and the FS-554-bound structures of *Lm*TryS (green) and *Li*TryS (dark green). The RMSD of the inhibitor FS-554 in complex with *Lm*TryS (yellow) and *Li*TryS (dark yellow) is additionally given.

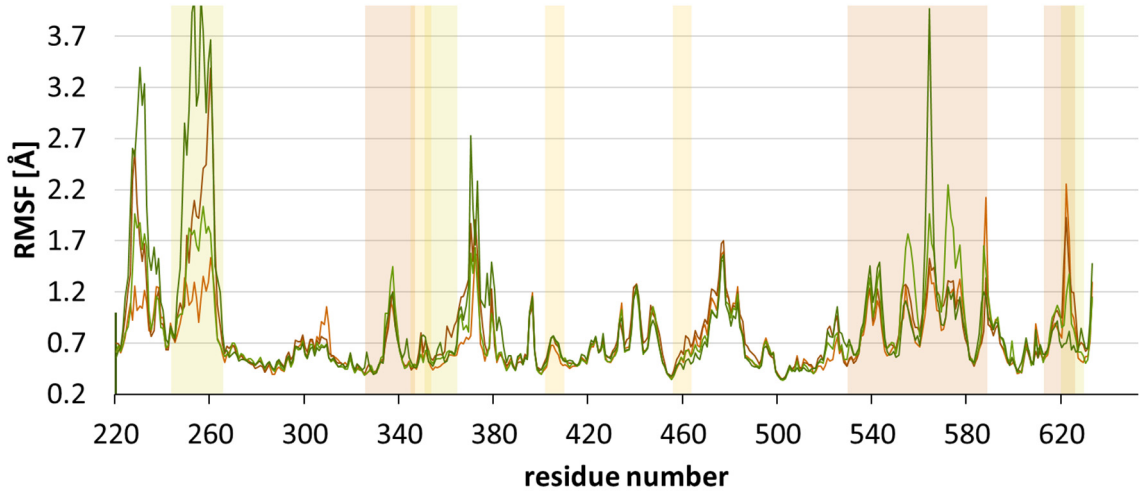


Figure A4.9. RMSF plots for different MD simulations of *Lm*TryS and *Li*TryS in complex with FS-554. The RMSF of the residues' Cα atoms is given for the 600 ns MD simulations of the ligand-free structures of *Lm*TryS (red) and *Li*TryS (dark red), and the FS-554-bound structures of *Lm*TryS (green) and *Li*TryS (dark green). The structures were aligned to the first MD snapshot with respect to the Cα atoms of resiudes 220 to 633 (residues 634 to 652 were excluded due to the high flexibility of the C-terminus in the simulations with inhibitor). Residues of the amidase domain, ATP-, GSH-, and GSP-binding residues are highlighted by green, red, yellow, and light green bars, respectively.

Table A4.3. Hydrogen bond analysis results for FS-554 in the MD simulations in complex with *Lm*TryS and *Li*TryS. Frequently observed hydrogen bonds for compound FS-554 in complex with TryS. Hydrogen bonding interactions which are observed in at least 5% of the 600 ns MD simulation are given. Equivalent residues in both enzymes after sequence alignment are given in the same row. The asterisk denotes mainchain hydrogen bond interactions.

***Lm*TryS (FS-554)**		***Li*TryS (FS-554)**	
hydrogen bond partner	percentage	hydrogen bond partner	percentage
Phe586*	60.17%	Phe586*	58.00%
Glu587	54.67%	Glu587	6.17%
Val618*	14.82%	Val618*	18.17%
Ile619*	13.17%	Ile619*	14.83%
Gln583	11.83%	Gln583	5.17%
Asp330	*5.33%*		
		*Leu588**	*8.83%*
		*Gly555**	*6.50%*
		*Gln584**	*13.00%*

Table A4.4. The ATP binding site residues of *Lm*TryS, *Li*TryS, *Tb*TryS, and *Tc*TryS after an structure-based alignment of the homology models. The numbering corresponds to the residues of *Lm*TryS in the structure with the PDB-ID 2vpm. The asterisks mark positions of the enzymes that harbour different residues.

	328	330	343	344	346	513	530	546	548	552	553	554	555	558	560	583	584	585	586	587	588	595	596	597	618	619	620	621	622
								*							*		*			*	*				*	*			*
***Lm*TryS**	R	D	F	E	N	K	L	A	K	G	R	V	G	V	I	Q	Q	L	F	E	L	Y	A	I	V	I	T	G	V
***Li*TryS**	R	D	F	E	N	K	L	A	K	G	R	V	G	V	I	Q	Q	L	F	E	L	Y	A	I	V	I	T	G	V
***Tb*TryS**	R	D	F	E	N	K	L	A	K	G	R	V	G	V	V	Q	Q	L	F	R	I	Y	A	I	L	T	T	G	L
***Tc*TryS**	R	D	F	E	N	K	L	V	K	G	R	V	G	V	V	Q	E	L	F	R	L	Y	A	I	I	I	T	G	L

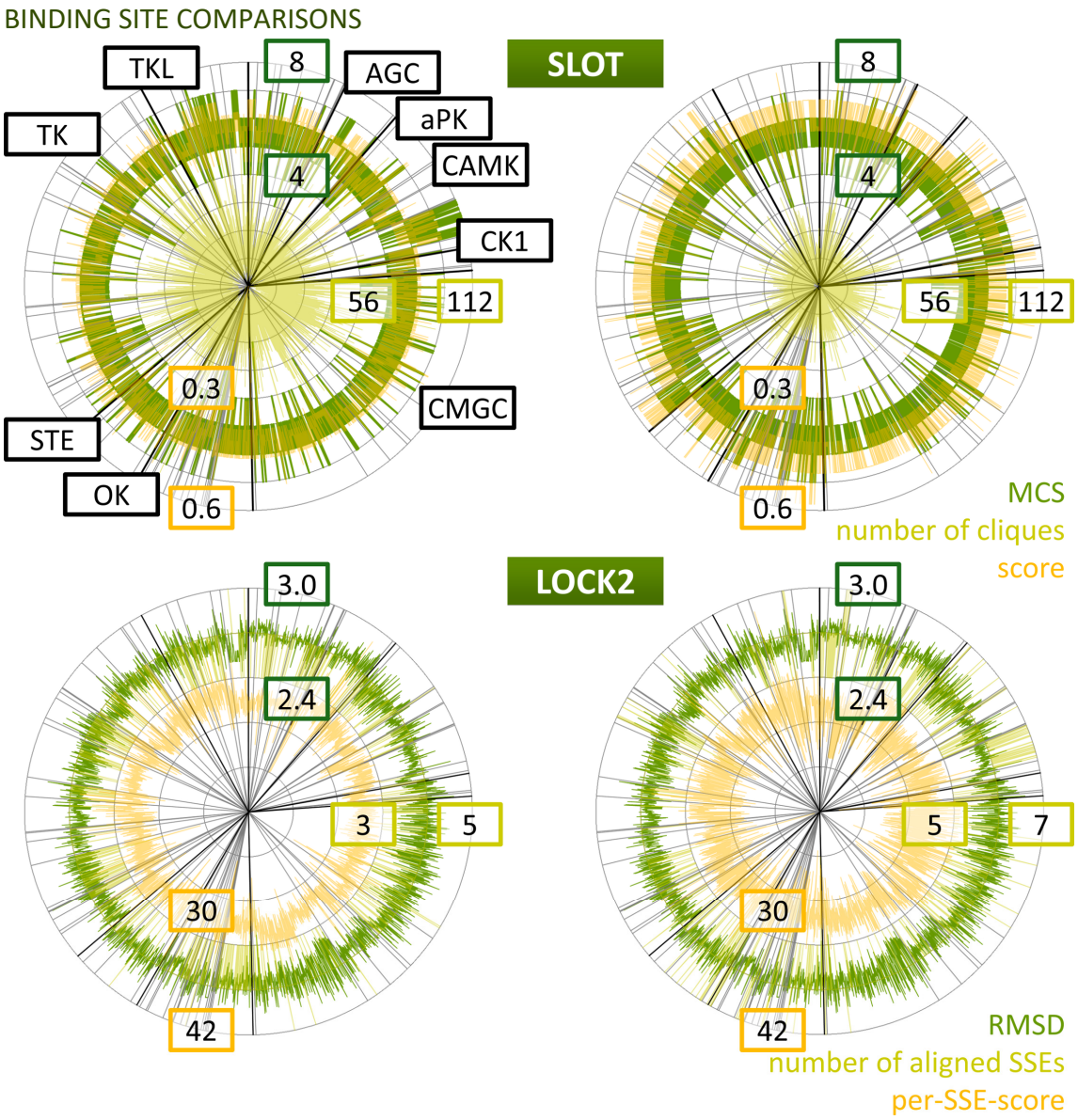


Figure A4.10. Secondary structure similarities between the ATP binding sites of TryS and protein kinases. The radar plots of MCS (number of matched SSE elements), number of cliques, and score for the SLOT-derived matches (top) and RMSD, number of aligned SSEs, and per-SSE-scores for the LOCK2 alignments (bottom) of the excised ATP binding sites of TryS and protein kinases of the KLIFS database are given. The results for the comparison between *Lm*TryS and *Li*TryS representative structures are given on the left and right, respectively. The kinase structures are grouped according to their membership to protein kinase groups as defined by Manning and co-workers[3]: AGC (protein kinase families A, G, and C), aPK (atypical protein kinases), CAMK (Ca^{2+}/calmodulin-dependent protein kinase), CK1 (casein kinase 1), CMGC (group including CDK, MAPK, GSK, and CDK-like kinase families), OK (other kinases including e.g. Aurora kinases and CDC7), TK (tyrosine kinase), and TKL (tyrosine kinase-like).

Table A4.5. Matches between *Li*TryS and protein kinases with a consensus of at least five methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Li*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
4nif.B	ANP	Erk2	MAPK	CMGC	li01A	1.03	0.45	0.19	IsoMIF, ProBiS, SiteHopper, SMAP, TM-align	5
4h3q.A	ANP	Erk2	MAPK	CMGC	li01A	0.99	0.47	0.17	IsoMIF, ProBiS, SiteAlign, SiteHopper, SMAP	5
3tei.A	ANP	Erk2	MAPK	CMGC	li00A	0.99	0.39	0.20	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3p0m.A	4SB	PKACa	PKA	AGC	li00A	0.96	0.38	0.20	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4dn5.B	AGS	NIK	STE-Unique	STE	li01A	0.95	0.39	0.19	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4fl2.A	ANP	SYK	Syk	TK	li01A	0.95	0.38	0.19	IsoMIF, SiteEngine, SiteHopper, SMAP, TM-align	5
3p23.C	ADP	IRE1	IRE	Other	li01A	0.95	0.44	0.17	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4iz5.A	ADP	Erk2	MAPK	CMGC	li00A	0.94	0.37	0.19	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3e64.A	5B3	JAK2	JakA	TK	li00A	0.94	0.34	0.20	IsoMIF, ProBiS, SiteAlign, SiteHopper, SMAP	5
3eqd.A	AGS	MAP2K1	STE7	STE	li01A	0.94	0.41	0.18	Cavbase, IsoMIF, ProBiS, SiteHopper, SMAP	5
4u40.A	ANP	HGK	STE20	STE	li01A	0.94	0.42	0.18	Cavbase, IsoMIF, ProBiS, SiteHopper, SMAP	5
5dnr.A	ATP	AurA	AUR	Other	li01A	0.94	0.41	0.18	Cavbase, IsoMIF, SiteEngine, SiteHopper, SMAP	5
3dfc.B	ANP	DAPK1	DAPK	CAMK	li02A	0.94	0.42	0.17	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
5d41.B	ANP	EGFR	EGFR	TK	li01A	0.93	0.40	0.18	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4iz5.D	ADP	Erk2	MAPK	CMGC	li03A	0.93	0.35	0.19	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
4ceg.A	ADP	AurA	AUR	Other	li01A	0.93	0.39	0.18	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
2oid.D	ANP	IRAK4	IRAK	TKL	li01A	0.93	0.39	0.18	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5

Table A4.5 *(continued)*. Matches between *Li*TryS and protein kinases with a consensus of at least five methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Li*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
2a19.B	ANP	PKR	PEK	Other	li00A	0.93	0.36	0.19	IsoMIF, Shaper, SiteEngine, SiteHopper, SMAP	5
2cci.A	ATP	CDK2	CDK	CMGC	li00A	0.93	0.31	0.20	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4h3p.A	ANP	Erk2	MAPK	CMGC	li00A	0.92	0.36	0.19	IsoMIF, SiteAlign, SiteEngine, SiteHopper, SMAP	5
2ozo.A	ANP	ZAP70	Syk	TK	li01A	0.91	0.41	0.17	IsoMIF, SiteAlign, SiteEngine, SiteHopper, SMAP	5
5cno.A	ANP	EGFR	EGFR	TK	li01A	0.90	0.37	0.18	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3gu4.A	ANP	DAPK1	DAPK	CAMK	li01A	0.90	0.42	0.16	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper	5
5v60.A	ACP	Erk2	MAPK	CMGC	li00A	0.90	0.36	0.18	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3p23.D	ADP	IRE1	IRE	Other	li01A	0.90	0.40	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
5av0.A	47X	DAPK1	DAPK	CAMK	li01A	0.89	0.35	0.18	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3e92.A	G6A	p38a	MAPK	CMGC	li03A	0.89	0.40	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3gu6.A	ADP	DAPK1	DAPK	CAMK	li02A	0.89	0.39	0.16	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
4nst.A	ADP	CRK7	CDK	CMGC	li00A	0.89	0.37	0.17	IsoMIF, ProBiS, SiteHopper, SMAP, TM-align	5
1rqq.A	112	INSR	InsR	TK	li00A	0.88	0.35	0.18	IsoMIF, SiteAlign, SiteEngine, SiteHopper, SMAP	5
4ic7.D	ANP	Erk5	MAPK	CMGC	li00A	0.88	0.40	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3gu5.A	ANP	DAPK1	DAPK	CAMK	li01A	0.88	0.39	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3eh9.A	ADP	DAPK1	DAPK	CAMK	li02A	0.87	0.37	0.17	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4f9a.A	ADP	CDC7	CDC7	Other	li00A	0.87	0.36	0.17	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5

Table A4.5 *(continued)*. Matches between *Li*TryS and protein kinases with a consensus of at least five methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Li*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
1ua2.B	ATP	CDK7	CDK	CMGC	li01A	0.87	0.33	0.18	IsoMIF, SiteEngine, SiteHopper, SMAP, TM-align	5
3eqb.A	ATP	MAP2K1	STE7	STE	li01A	0.87	0.37	0.17	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3sls.B	ANP	MAP2K2	STE7	STE	li01A	0.86	0.39	0.16	IsoMIF, ProBiS, SiteAlign, SiteHopper, SMAP	5
1yxt.A	ANP	PIM1	PIM	CAMK	li01A	0.86	0.44	0.14	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP, TM-align	6
4cxa.A	ANP	CRK7	CDK	CMGC	li02A	0.86	0.39	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP, TM-align	6
4zse.B	ANP	EGFR	EGFR	TK	li02A	0.86	0.37	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP, TM-align	6
5k4i.A	6QB	Erk2	MAPK	CMGC	li00A	0.86	0.34	0.17	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
5efq.A	ADP	CHED	CDK	CMGC	li02A	0.86	0.38	0.16	IsoMIF, ProBiS, SiteHopper, SMAP, TM-align	5
3orn.A	ANP	MAP2K1	STE7	STE	li01A	0.85	0.36	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4u80.A	ANP	MAP2K1	STE7	STE	li01A	0.85	0.36	0.17	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3a7j.A	ADP	MST3	STE20	STE	li01A	0.85	0.38	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3dv3.A	ATP	MAP2K1	STE7	STE	li01A	0.85	0.34	0.17	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3s4q.A	NK0	p38a	MAPK	CMGC	li03A	0.85	0.37	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4mne.H	ACP	MAP2K1	STE7	STE	li02A	0.85	0.39	0.15	IsoMIF, ProBiS, SiteAlign, SiteHopper, SMAP	5
1s9i.B	ATP	MAP2K2	STE7	STE	li01A	0.85	0.35	0.17	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
5ng0.A	ACP	RIPK2	RIPK	TKL	li00A	0.84	0.30	0.18	Cavbase, IsoMIF, ProBiS, SiteHopper, SMAP	5
5kup.A	6XL	BTK	Tec	TK	li00A	0.84	0.37	0.16	IsoMIF, ProBiS, SiteHopper, SMAP, TM-align	5

Table A4.5 *(continued)*. Matches between *Li*TryS and protein kinases with a consensus of at least five methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Li*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
2w4k.A	ADP	DAPK1	DAPK	CAMK	li02A	0.84	0.36	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4l6q.B	1WU	ROCK2	DMPK	AGC	li00A	0.84	0.37	0.16	Cavbase, IsoMIF, ProBiS, SiteHopper, SMAP	5
3vvh.B	ATP	MAP2K1	STE7	STE	li01A	0.84	0.36	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
4nst.C	ADP	CRK7	CDK	CMGC	li01A	0.84	0.38	0.15	IsoMIF, SiteEngine, SiteHopper, SMAP, TM-align	5
5ukl.A	SIX	BARK1	GRK	AGC	li02A	0.84	0.41	0.14	IsoMIF, Shaper, SiteEngine, SiteHopper, SMAP	5
3d7z.A	GK5	p38a	MAPK	CMGC	li03A	0.84	0.35	0.16	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
4xlv.A	ACP	INSR	InsR	TK	li00A	0.83	0.37	0.15	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3bx5.A	304	p38a	MAPK	CMGC	li03A	0.83	0.38	0.15	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
1ig1.A	ANP	DAPK1	DAPK	CAMK	li00A	0.82	0.34	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
5cal.A	4Z8	EGFR	EGFR	TK	li03A	0.82	0.34	0.16	ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	5
4aac.A	AAV	p38a	MAPK	CMGC	li03A	0.82	0.38	0.15	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	5
3uoh.A	0C4	AurA	AUR	Other	li01A	0.82	0.36	0.15	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP, TM-align	6
5efq.C	ADP	CHED	CDK	CMGC	li01A	0.82	0.34	0.16	IsoMIF, SiteEngine, SiteHopper, SMAP, TM-align	5
3que.A	3FF	p38a	MAPK	CMGC	li00A	0.81	0.36	0.15	Cavbase, ProBiS, SiteEngine, SMAP, TM-align	5
2w5a.A	ADP	NEK2	NEK	Other	li02A	0.80	0.35	0.15	Cavbase, IsoMIF, SiteEngine, SMAP, TM-align	5
3twj.C	07R	ROCK1	DMPK	AGC	li01A	0.79	0.35	0.15	IsoMIF, ProBiS, SiteEngine, SMAP, TM-align	5
3uoh.B	0C4	AurA	AUR	Other	li01A	0.77	0.37	0.13	IsoMIF, ProBiS, SiteEngine, SMAP, TM-align	5

Table A4.5 *(continued)*. Matches between *Li*TryS and protein kinases with a consensus of at least five methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Li*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
1wbv.A	LI3	p38a	MAPK	CMGC	li01A	0.72	0.34	0.13	Cavbase, ProBiS, Shaper, SiteEngine, SMAP	5
3o8p.A	BMU	p38a	MAPK	CMGC	li01A	0.70	0.34	0.12	Cavbase, IsoMIF, ProBiS, SiteEngine, SMAP	5

Table A4.6. Matches between *Lm*TryS and protein kinases with a consensus of at least six methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Lm*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
4ark.A	ADP	MAP2K1	STE7	STE	lm01A	1.03	0.39	0.21	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteHopper, SMAP	6
4wb5.A	ATP	PKACa	PKA	AGC	lm01A	1.02	0.45	0.19	Cavbase, IsoMIF, ProBiS, Shaper, SiteHopper, SMAP	6
4iz5.A	ADP	Erk2	MAPK	CMGC	lm02A	0.96	0.36	0.20	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4nif.B	ANP	Erk2	MAPK	CMGC	lm00A	0.96	0.41	0.18	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	7
3eqd.A	AGS	MAP2K1	STE7	STE	lm01A	0.95	0.39	0.19	Cavbase, IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	7
3gu5.A	ANP	DAPK1	DAPK	CAMK	lm00A	0.95	0.42	0.18	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4zse.A	ANP	EGFR	EGFR	TK	lm01A	0.94	0.37	0.19	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4fv5.A	EK9	Erk2	MAPK	CMGC	lm00A	0.94	0.38	0.19	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
4cxa.A	ANP	CRK7	CDK	CMGC	lm01A	0.94	0.40	0.18	Cavbase, IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	7
3dfc.B	ANP	DAPK1	DAPK	CAMK	lm00A	0.94	0.42	0.17	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
3f5u.A	ANP	DAPK1	DAPK	CAMK	lm00A	0.94	0.43	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4h3q.A	ANP	Erk2	MAPK	CMGC	lm00A	0.94	0.40	0.18	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
3gu4.A	ANP	DAPK1	DAPK	CAMK	lm00A	0.93	0.44	0.17	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
5aep.A	QUP	JAK2	JakA	TK	lm00A	0.93	0.35	0.19	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
4fv6.A	E57	Erk2	MAPK	CMGC	lm03A	0.93	0.37	0.19	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6

Table A4.6 *(continued)*. Matches between *Lm*TryS and protein kinases with a consensus of at least six methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Lm*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
4f99.A	ADP	CDC7	CDC7	Other	lm01A	0.93	0.34	0.20	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3i60.A	E86	Erk2	MAPK	CMGC	lm03A	0.93	0.36	0.19	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
3ncz.B	3NC	ROCK1	DMPK	AGC	lm01A	0.93	0.38	0.18	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
3gu7.A	ADP	DAPK1	DAPK	CAMK	lm02A	0.93	0.39	0.18	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
1ol5.A	ADP	AurA	AUR	Other	lm07A	0.92	0.38	0.18	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3dv3.A	ATP	MAP2K1	STE7	STE	lm01A	0.91	0.35	0.19	IsoMIF, ProBiS, Shaper, SiteAlign, SiteEngine, SiteHopper, SMAP	7
3dy7.A	ATP	MAP2K1	STE7	STE	lm01A	0.91	0.33	0.19	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteHopper, SMAP	6
3pp1.A	ATP	MAP2K1	STE7	STE	lm01A	0.91	0.39	0.17	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
1ua2.D	ATP	CDK7	CDK	CMGC	lm01A	0.91	0.40	0.17	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP, TMalign	6
4ic7.D	ANP	Erk5	MAPK	CMGC	lm03A	0.91	0.39	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4ekk.A	ANP	AKT1	Akt	AGC	lm01A	0.90	0.34	0.19	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
3u87.B	ANP	CK2a1	CK2	CMGC	lm02A	0.90	0.36	0.18	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper	6
5nhh.A	8XH	Erk2	MAPK	CMGC	lm03A	0.90	0.38	0.17	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	7
3mbl.A	ADP	MAP2K1	STE7	STE	lm01A	0.89	0.37	0.18	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
1qcf.A	PP1	HCK	Src	TK	lm02A	0.89	0.36	0.18	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6

Table A4.6 *(continued)*. Matches between *Lm*TryS and protein kinases with a consensus of at least six methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Lm*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
3gu6.A	ADP	DAPK1	DAPK	CAMK	lm00A	0.89	0.38	0.17	IsoMIF, ProBiS, Shaper, SiteAlign, SiteEngine, SiteHopper, SMAP	7
3mv5.A	XFE	AKT1	Akt	AGC	lm06A	0.89	0.31	0.19	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
5d41.A	ANP	EGFR	EGFR	TK	lm01A	0.89	0.36	0.18	Cavbase, IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	7
3twj.C	07R	ROCK1	DMPK	AGC	lm02A	0.88	0.35	0.18	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP, TMalign	6
2b7a.B	IZA	JAK2	JakA	TK	lm00A	0.88	0.36	0.17	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
4zse.B	ANP	EGFR	EGFR	TK	lm01A	0.88	0.34	0.18	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
4hvi.A	19S	JAK3	JakA	TK	lm03A	0.87	0.34	0.18	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3lmg.B	ANP	ErbB3	EGFR	TK	lm01A	0.87	0.31	0.19	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	7
4fg7.A	ATP	CaMK1a	CAMK1	CAMK	lm03A	0.87	0.38	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4riw.D	ADP	EGFR	EGFR	TK	lm03A	0.87	0.36	0.17	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
4riy.D	ADP	EGFR	EGFR	TK	lm03A	0.87	0.35	0.17	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
4oav.D	ACP	RNAseL	Other-Unique	Other	lm01A	0.87	0.30	0.19	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
5uiq.C	8BV	IRAK4	IRAK	TKL	lm01A	0.87	0.31	0.19	IsoMIF, Shaper, SiteAlign, SiteEngine, SiteHopper, SMAP	6
2cci.C	ATP	CDK2	CDK	CMGC	lm00A	0.87	0.32	0.18	Cavbase, IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	7

Table A4.6 *(continued)*. Matches between *Lm*TryS and protein kinases with a consensus of at least six methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Lm*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
4fl3.A	ANP	SYK	Syk	TK	lm01A	0.87	0.35	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4qp8.B	362	Erk2	MAPK	CMGC	lm01A	0.87	0.40	0.16	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
3aln.B	ANP	MAP2K4	STE7	STE	lm00A	0.87	0.36	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3dls.C	ADP	PASK	CAMKL	CAMK	lm01A	0.86	0.38	0.16	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
3eh9.A	ADP	DAPK1	DAPK	CAMK	lm00A	0.86	0.36	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3alo.A	ANP	MAP2K4	STE7	STE	lm02A	0.86	0.33	0.18	IsoMIF, ProBiS, Shaper, SiteAlign, SiteEngine, SiteHopper, SMAP	7
3aln.A	ANP	MAP2K4	STE7	STE	lm00A	0.86	0.38	0.16	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4oti.A	MI1	PKN1	PKN	AGC	lm01A	0.86	0.34	0.17	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	7
4fl2.A	ANP	SYK	Syk	TK	lm02A	0.86	0.41	0.15	IsoMIF, ProBiS, Shaper, SiteHopper, SMAP, TMalign	6
4rix.A	ANP	ErbB3	EGFR	TK	lm01A	0.85	0.30	0.18	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4yve.B	4KK	ROCK1	DMPK	AGC	lm01A	0.85	0.32	0.18	Cavbase, IsoMIF, ProBiS, Shaper, SiteHopper, SMAP	6
4bie.A	IE6	MAP3K5	STE11	STE	lm07A	0.85	0.37	0.16	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3bu5.A	ATP	INSR	InsR	TK	lm04A	0.85	0.38	0.16	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	7
2cn5.A	ADP	CHK2	RAD53	CAMK	lm03A	0.85	0.38	0.16	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6

Table A4.6 *(continued)*. Matches between *Lm*TryS and protein kinases with a consensus of at least six methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Lm*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
5c03.A	AGS	TYK2-b	JakB	TK	lm01A	0.85	0.36	0.16	IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP, TMalign	6
5k5n.A	6QH	GSK3B	GSK	CMGC	lm01A	0.84	0.33	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
4qml.A	ANP	MST3	STE20	STE	lm02A	0.84	0.28	0.19	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
2yix.A	YIX	p38a	MAPK	CMGC	lm02A	0.84	0.42	0.14	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
3zxt.B	ACP	DAPK1	DAPK	CAMK	lm02A	0.84	0.33	0.17	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
2w4k.A	ADP	DAPK1	DAPK	CAMK	lm07A	0.84	0.37	0.16	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3v5w.A	8PR	BARK1	GRK	AGC	lm00A	0.84	0.33	0.17	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
4eh3.A	NAR	p38a	MAPK	CMGC	lm01A	0.84	0.42	0.14	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
3eqg.A	ADP	MAP2K1	STE7	STE	lm00A	0.84	0.33	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
5bvf.A	4VJ	Erk2	MAPK	CMGC	lm03A	0.84	0.35	0.16	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
3vqh.A	IQB	PKACa	PKA	AGC	lm01A	0.84	0.40	0.15	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
4rix.D	ADP	EGFR	EGFR	TK	lm03A	0.83	0.37	0.16	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
3orn.A	ANP	MAP2K1	STE7	STE	lm01A	0.83	0.36	0.16	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
2ydk.A	YDK	CHK1	CAMKL	CAMK	lm03A	0.83	0.32	0.17	Cavbase, IsoMIF, ProBiS, SiteEngine, SiteHopper, SMAP	6
2xuu.A	ADP	DAPK1	DAPK	CAMK	lm07A	0.83	0.36	0.16	IsoMIF, ProBiS, Shaper, SiteAlign, SiteEngine, SiteHopper, SMAP	7

Table A4.6 *(continued)*. Matches between *Lm*TryS and protein kinases with a consensus of at least six methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Lm*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
4h3p.A	ANP	Erk2	MAPK	CMGC	lm03A	0.83	0.33	0.17	Cavbase, IsoMIF, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3tjd.B	6TP	JAK2	JakA	TK	lm00A	0.83	0.35	0.16	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper	6
2a19.B	ANP	PKR	PEK	Other	lm02A	0.83	0.34	0.16	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
1ua2.B	ATP	CDK7	CDK	CMGC	lm01A	0.82	0.36	0.16	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
3dls.A	ADP	PASK	CAMKL	CAMK	lm01A	0.82	0.37	0.15	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
4rix.C	ANP	ErbB3	EGFR	TK	lm01A	0.82	0.31	0.17	IsoMIF, ProBiS, SiteAlign, SiteEngine, SiteHopper, SMAP	6
2y4i.B	ATP	KSR2	RAF	TKL	lm01A	0.82	0.33	0.16	Cavbase, IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	7
3soc.B	GVD	ACTR2	STKR	TKL	lm02A	0.82	0.38	0.15	IsoMIF, ProBiS, Shaper, SiteEngine, SiteHopper, SMAP	6
3nw6.A	LGW	IGF1R	InsR	TK	lm06A	0.82	0.33	0.16	Cavbase, IsoMIF, SiteAlign, SiteEngine, SMAP, TMalign	6
3vvh.C	ATP	MAP2K1	STE7	STE	lm01A	0.82	0.35	0.16	IsoMIF, ProBiS, Shaper, SiteAlign, SiteEngine, SMAP	6
4mne.D	ACP	MAP2K1	STE7	STE	lm01A	0.81	0.32	0.16	Cavbase, IsoMIF, ProBiS, Shaper, SiteEngine, SMAP	6
4an9.A	ACP	MAP2K1	STE7	STE	lm01A	0.81	0.29	0.17	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SMAP	6
2oji.A	33A	Erk2	MAPK	CMGC	lm08A	0.79	0.27	0.17	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SMAP	6
4fvq.A	ATP	JAK2-b	JakB	TK	lm01A	0.78	0.39	0.13	IsoMIF, ProBiS, Shaper, SiteEngine, SMAP, TMalign	6
1rqq.B	112	INSR	InsR	TK	lm02A	0.77	0.33	0.15	IsoMIF, ProBiS, Shaper, SiteAlign, SiteEngine, SMAP	6
3tv4.B	TV4	BRAF	RAF	TKL	lm00A	0.76	0.35	0.14	Cavbase, IsoMIF, ProBiS, Shaper, SiteEngine, SMAP	6

Table A4.6 *(continued)*. Matches between *Lm*TryS and protein kinases with a consensus of at least six methods. The kinase corresponding the the protein structure and the scores of a SiteHopper-based binding site alignment (PatchScore, ShapeTanimoto, ColorTanimoto)[4] are given. The comparison methods are listed together with the number of comparison approaches which identified the match (N). Erk2 and PKACA structures are highlighted in light and dark gray, respectively. A colour gradient ranging from green (maximum score) to yellow to red (minimum score) is applied for the PatchScore.

PDB-ID. Chain	Ligand ID	Kinase	Family	Group	*Lm*-TryS	Patch-Score	Shape-Tani-moto	Color-Tani-moto	Comparison Methods	N
4k18.A	1OB	PIM1	PIM	CAMK	lm07A	0.74	0.35	0.13	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SMAP	6
5khw.B	ADP	JAK1	JakA	TK	lm02A	0.74	0.30	0.14	IsoMIF, ProBiS, SiteAlign, SiteEngine, SMAP, TMalign	6
5ugx.A	ACP	FGFR2	FGFR	TK	lm03A	0.71	0.30	0.14	Cavbase, IsoMIF, ProBiS, SiteAlign, SiteEngine, SMAP	6

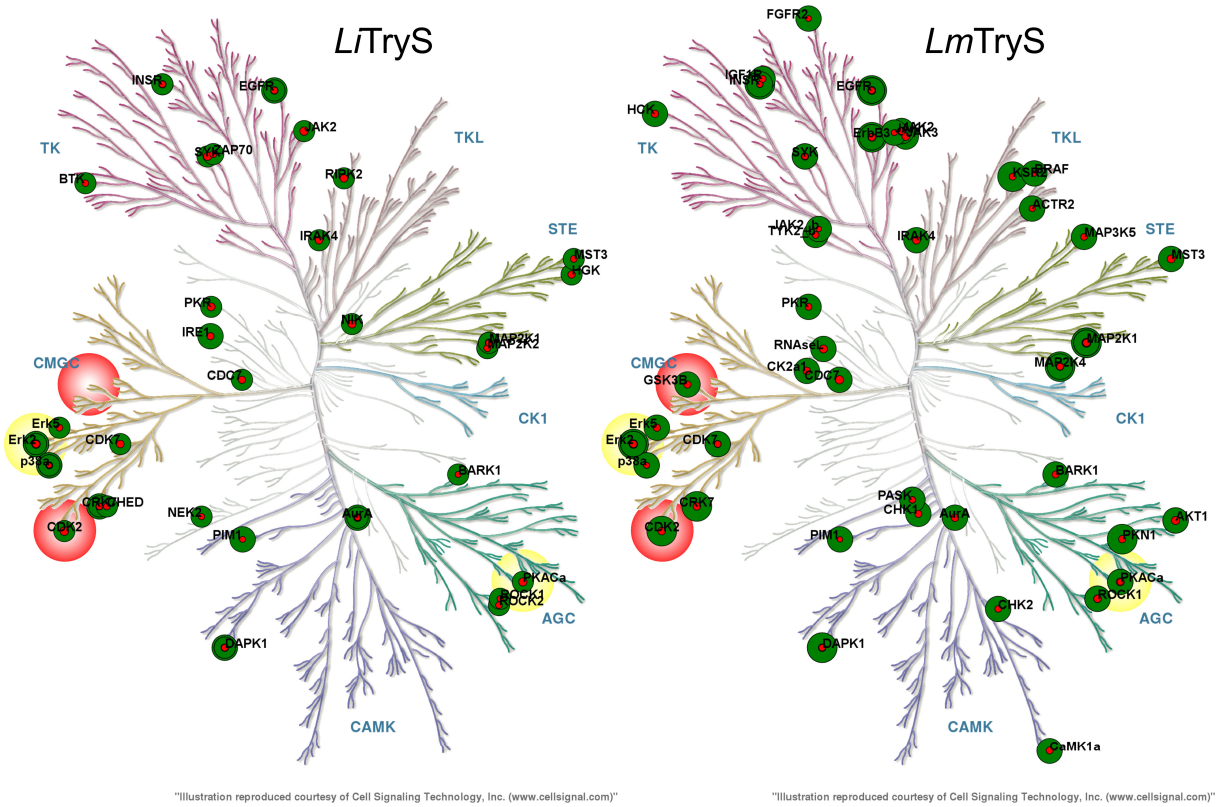


Figure A4.11. Kinome map showing protein kinases whose ATP binding site show high similarities to TryS. The circles highlight kinases whose binding sites were identified as similar to *Li*TryS (left) and *Lm*TryS (right) by at least five and six binding site comparison methods, respectively. The radius of the green circles scales with the SiteHopper-derived TanimotoColor score, while the size of the small red circles corresponds to the number of methods which identified similarities. Large red circles highlight GSK-3β and CDK2 structures which are known to bind compounds with a paullone-scaffold whereas the yellow circles highlight the most similar kinases (with respect to the SiteHopper PatchScore) to both, *Li*TryS and *Lm*TryS. The figure was generated using KinMap[5] and hhe kinase structures a grouped according to their membership to protein kinase groups as defined by Manning and co-workers[3]: AGC (protein kinase families A, G, and C), aPK (atypical protein kinases), CAMK (Ca^{2+}/calmodulin-dependent protein kinase), CK1 (casein kinase 1), CMGC (group including CDK, MAPK, GSK, and CDK-like kinase families), OK (other kinases including e.g. Aurora kinases and CDC7), TK (tyrosine kinase), and TKL (tyrosine kinase-like)..

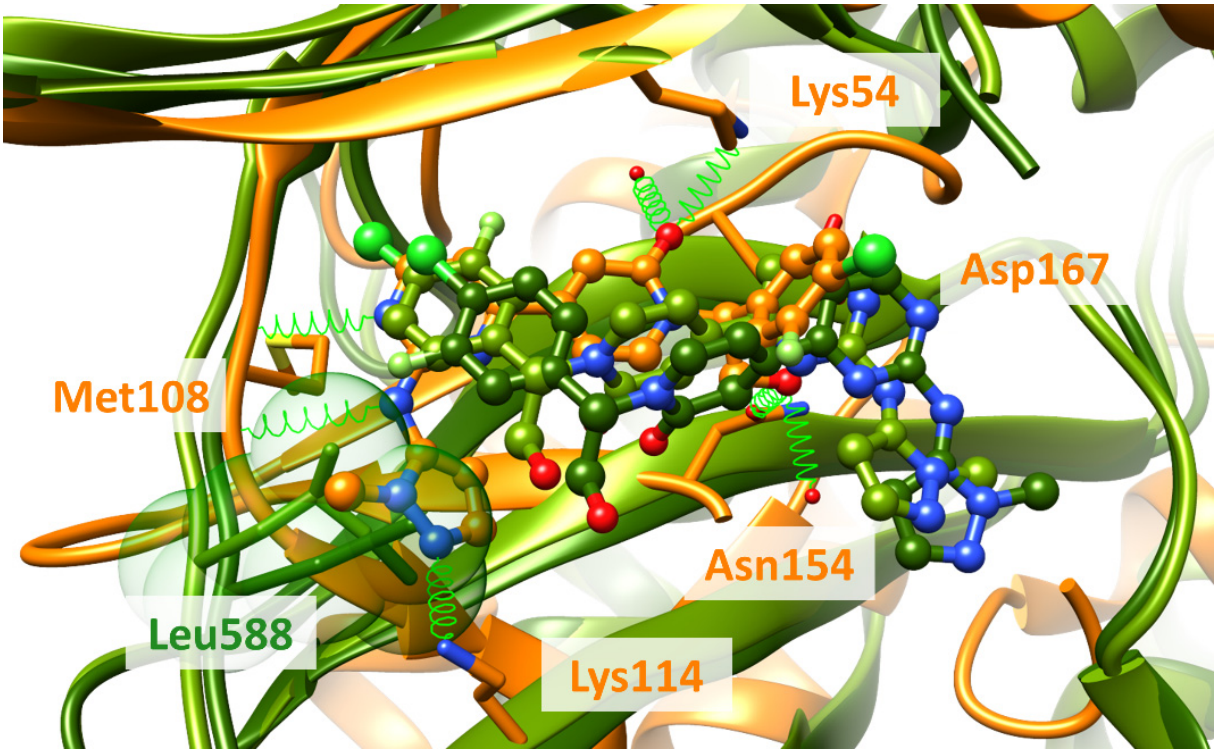


Figure A4.12. Binding modes of an Erk2 inhibitor in *Li*TryS and *Lm*TryS. The inhibitor with the PDB-ID 6QB of the original crystal structure in complex with Erk2 (orange, PDB-ID 5k4i) is presented together with its predicted binding modes in *Li*TryS (dark green) and *Lm*TryS (light green). The TryS structures are aligned to Erk2 based on the SiteHopper comparison. We observe a 180° horizontal rotation in the binding mode to TryS which can be attributed to the presence of residue Leu488 (semitransparent spheres) which blocks the part of the pocket which is occupied by the methylated pyrazole moiety in the Erk2 structure.

Table A4.7. Selected compounds after a docking of the Erk2 and PKACA inhibitor-based compound library in the ATP binding site of *Li*TryS. Structural parts depicted in green were predicted to bind to the front of the pocket via hydrogen bonds. Red moieties were predicted to bind in the region of the γ-phosphate in ATP. Yellow parts are situated in a hydrophobic sub-pocket in the respective docking poses. The substructures *N*-phenylbenzamide (a), diphenyl-methane (b), diphenylamine (c), diphenylsulfane (d), and 2-phenyl-4*H*-chromen-4-one (e) led to the identification of these compounds in the MolPort for docking purposes.

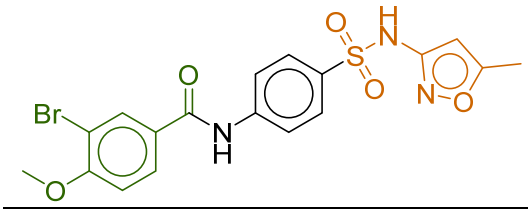
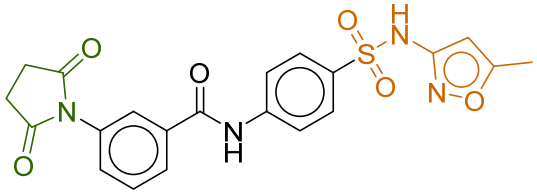
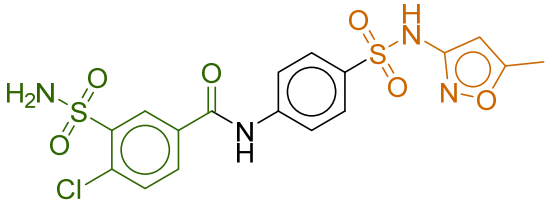
Structure	MolPort-ID	Name (ChEMBL Data)
	MolPort-001-516-942[a]	3-bromo-4-methoxy-*N*-(4-(*N*-(5-methylisoxazol-3-yl)-sulfamoyl)phenyl)-benzamide
	MolPort-000-681-826[a]	3-(2,5-dioxopyrrolidin-1-yl)-*N*-(4-(*N*-(5-methylisoxazol-3-yl)sulfamoyl)phenyl)-benzamide
	MolPort-005-850-488[a]	4-chloro-*N*-(4-(*N*-(5-methylisoxazol-3-yl)sulfamoyl)phenyl)-3-sulfamoylbenzamide

Table A4.7 *(continued)*. Selected compounds after a docking of the Erk2 and PKACA inhibitor-based compound library in the ATP binding site of *Li*TryS. Structural parts depicted in green were predicted to bind to the front of the pocket via hydrogen bonds. Red moieties were predicted to bind in the region of the γ-phosphate in ATP. Yellow parts are situated in a hydrophobic sub-pocket in the respective docking poses. The substructures *N*-phenylbenzamide (a), diphenylmethane (b), diphenylamine (c), diphenylsulfane (d), and 2-phenyl-4*H*-chromen-4-one (e) led to the identification of these compounds in the MolPort for docking purposes.

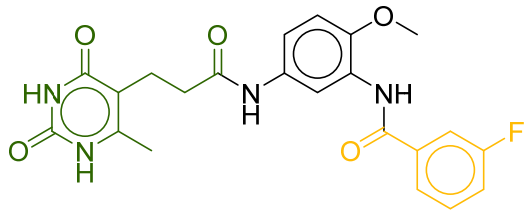
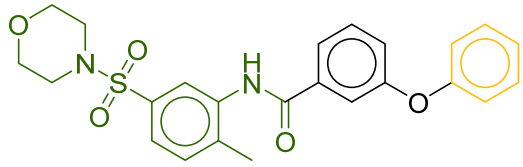
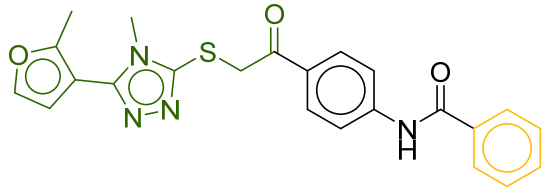
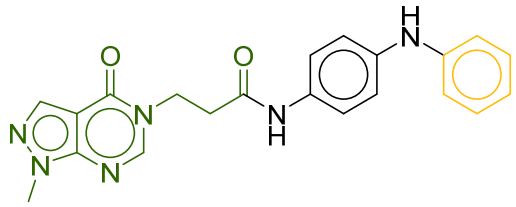
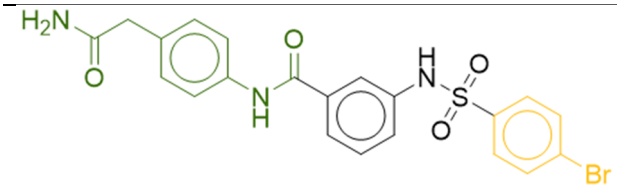
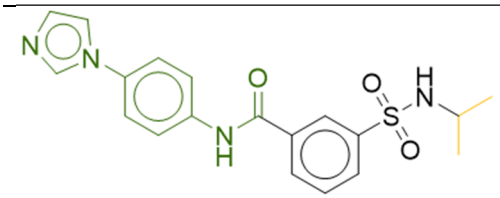
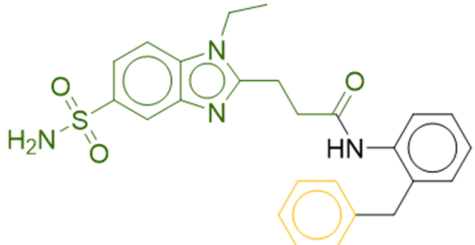
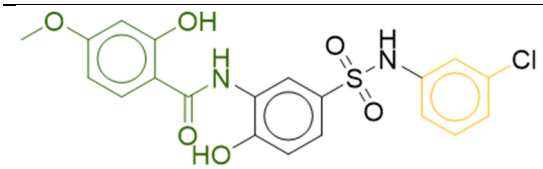
Structure	MolPort-ID	Name (ChEMBL Data)
	MolPort-007-528-273[a]	3-fluoro-*N*-(2-methoxy-5-(3-(6-methyl-2,4-dioxo-1,2,3,4-tetrahydropyrimidin-5-yl)propanamido)-phenyl)benzamide
	MolPort-005-578-527[a]	*N*-(2-methyl-5-(morpholinosulfonyl)-phenyl)-3-phenoxybenzamide
	MolPort-005-779-410[a]	*N*-(4-(2-((4-methyl-5-(2-methylfuran-3-yl)-4*H*-1,2,4-triazol-3-yl)thio)acetyl)-phenyl)benzamide
	MolPort-007-295-632[c]	3-(1-methyl-4-oxo-1,4-dihydro-5*H*-pyrazolo[3,4-*d*]pyrimidin-5-yl)-*N*-(4-(phenylamino)phenyl)-propanamide
	MolPort-009-187-150[a]	*N*-(4-(2-amino-2-oxoethyl)phenyl)-3-((4-bromophenyl)-sulfonamido)benzamide
	MolPort-009-169-553[a]	*N*-(4-(1*H*-imidazol-1-yl)phenyl)-3-(*N*-isopropylsulfamoyl)-benzamide
	MolPort-009-212-098[b]	*N*-(2-benzylphenyl)-3-(1-ethyl-5-sulfamoyl-1*H*-benzo[*d*]imidazol-2-yl)propanamide
	MolPort-009-253-334[a]	*N*-(5-(*N*-(3-chlorophenyl)sulfamoyl)-2-hydroxyphenyl)-2-hydroxy-4-methoxybenzamide

Table A4.7 *(continued)*. Selected compounds after a docking of the Erk2 and PKACA inhibitor-based compound library in the ATP binding site of *Li*TryS. Structural parts depicted in green were predicted to bind to the front of the pocket via hydrogen bonds. Red moieties were predicted to bind in the region of the γ-phosphate in ATP. Yellow parts are situated in a hydrophobic sub-pocket in the respective docking poses. The substructures *N*-phenylbenzamide (a), diphenylmethane (b), diphenylamine (c), diphenylsulfane (d), and 2-phenyl-4*H*-chromen-4-one (e) led to the identification of these compounds in the MolPort for docking purposes.

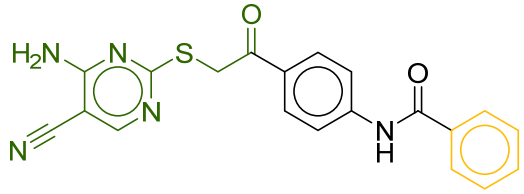
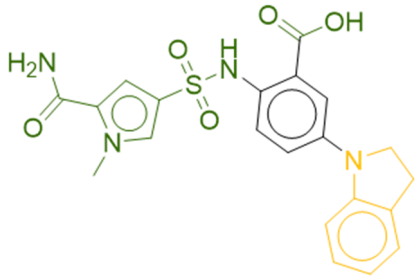
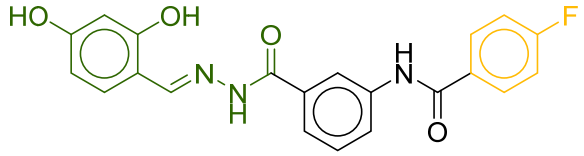
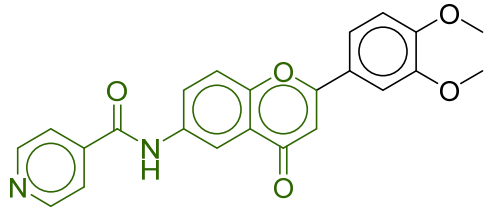
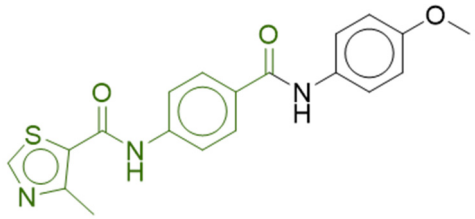
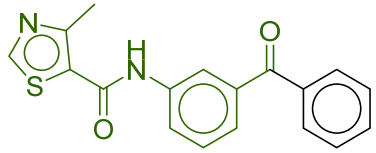
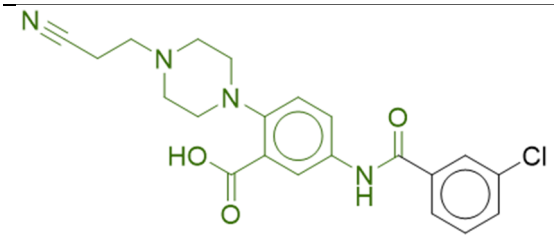
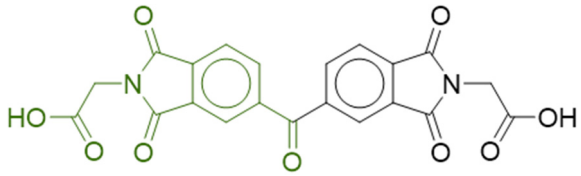
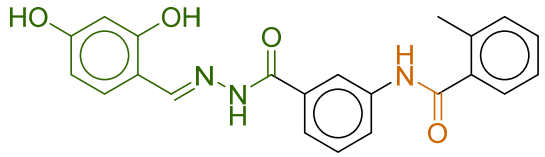
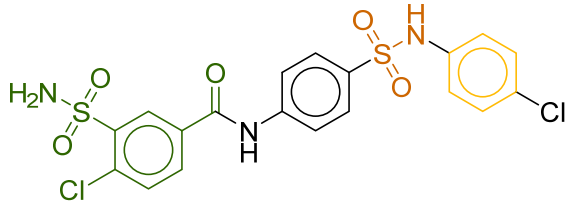
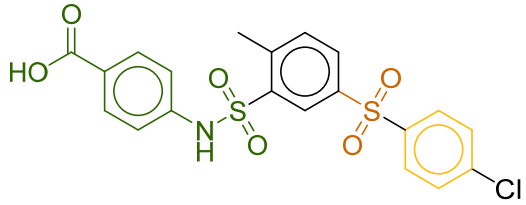
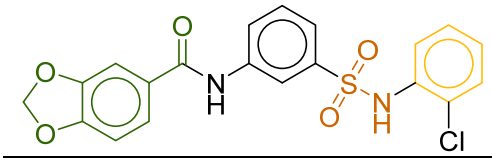
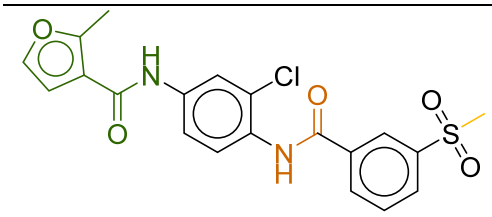
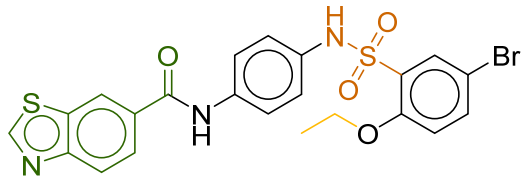
Structure	MolPort-ID	Name (ChEMBL Data)
	MolPort-019-649-259[a]	*N*-(4-(2-((4-amino-5-cyanopyrimidin-2-yl)thio)acetyl)phenyl)-benzamide
	MolPort-027-929-080[c]	2-((5-carbamoyl-1-methyl-1*H*-pyrrole)-3-sulfonamido)-5-(indolin-1-yl)benzoic acid
	MolPort-030-082-298[a]	(*E*)-*N*-(3-(2-(2,4-dihydroxybenzylidene)-hydrazine-1-carbonyl)phenyl)-4-fluorobenzamide
	MolPort-005-931-839[e]	*N*-(2-(3,4-dimethoxyphenyl)-4-oxo-4*H*-chromen-6-yl)isonicotinamide
	MolPort-009-330-431[a]	*N*-(4-((4-methoxyphenyl)-carbamoyl)phenyl)-4-methylthiazole-5-carboxamide (CHEMBL1889695, IC_{50}(HKDC1) > 80 µM)
	MolPort-019-678-765[b]	*N*-(3-benzoylphenyl)-4-methylthiazole-5-carboxamide
	MolPort-007-891-804[a]	5-(3-chlorobenzamido)-2-(4-(2-cyanoethyl)piperazin-1-yl)benzoic acid

Table A4.8. Selected compounds after a docking of the Erk2 and PKACA inhibitor-based compound library in the ATP binding site of *Lm*TryS. Structural parts depicted in green were predicted to bind to the front of the pocket via hydrogen bonds. Red moieties were predicted to bind in the region of the γ-phosphate in ATP. Yellow parts are situated in a hydrophobic sub-pocket in the respective docking poses. The asterisk highlights the compound whose binding mode is presented in the manuscript. The substructures *N*-phenylbenzamide (a), diphenyl-methane (b), diphenylamine (c), diphenylsulfane (d), and 2-phenyl-4*H*-chromen-4-one (e) led to the identification of these compounds in the MolPort for docking purposes.

Structure	MolPort-ID	Name
	MolPort-001-018-951[b]	2,2'-(carbonylbis(1,3-dioxoisoindoline-5,2-diyl))diacetic acid
	MolPort-007-553-340[a]	(*E*)-*N*-(3-(2-(2,4-dihydroxybenzylidene)-hydrazine-1-carbonyl)phenyl)-2-methylbenzamide
 *	MolPort-005-832-541[a]	4-chloro-*N*-(4-(*N*-(4-chlorophenyl)sulfamoyl) phenyl)-3-sulfamoyl-benzamide
	MolPort-001-893-271[d]	4-((5-((4-chlorophenyl)sulfonyl)-2-methylphenyl)-sulfonamido)benzoic acid
	MolPort-004-073-334[a]	*N*-(3-(*N*-(2-chlorophenyl)sulfamoyl)-phenyl)benzo[*d*][1,3]-dioxole-5-carboxamide
	MolPort-004-644-792[a]	*N*-(3-chloro-4-(3-(methylsulfonyl)-benzamido)phenyl)-2-methylfuran-3-carboxamide
	MolPort-005-866-484[a]	*N*-(4-((5-bromo-2-ethoxyphenyl)-sulfonamido)phenyl)-benzo[*d*]thiazole-6-carboxamide

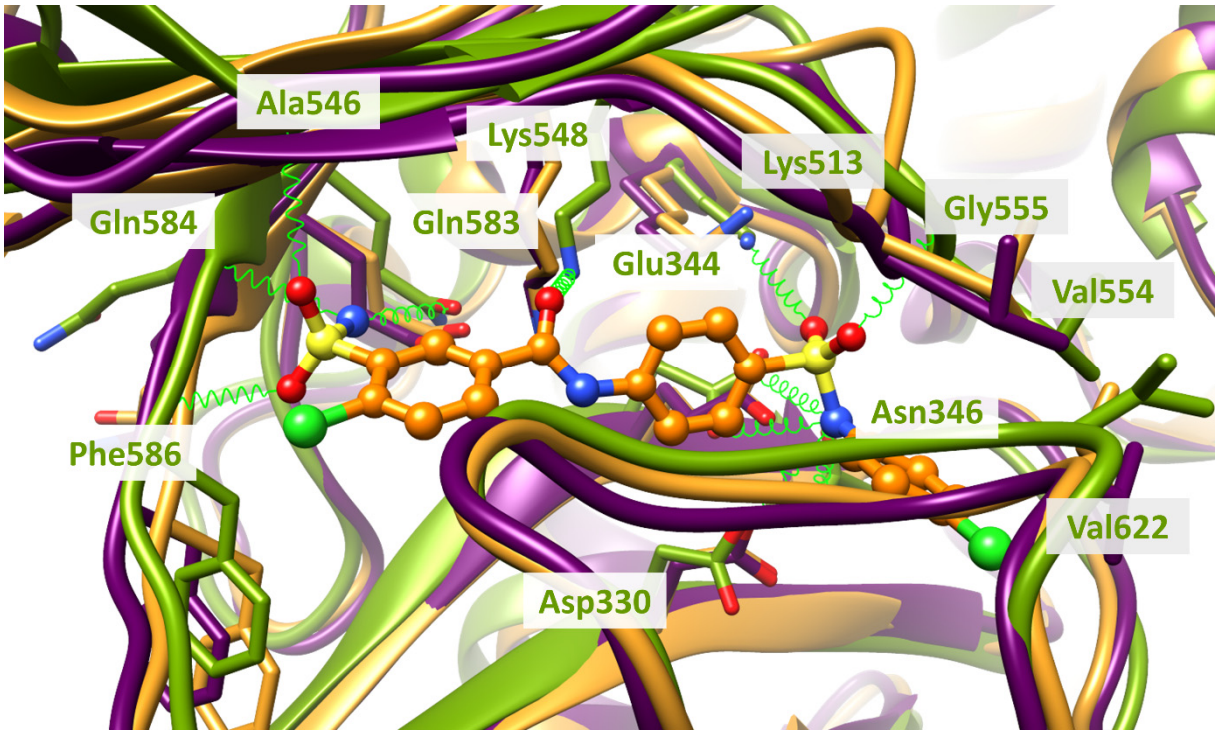


Figure A4.13. Alignment of *Tb*TryS and *Tc*TryS to *Lm*TryS in complex with a promising virtual hit compound. The compound with the MolPort-ID 005-832-541 (orange, ball-and-stick representation) interacts with *Lm*TryS (green cartoon representation) via many hydrogen bond interactions (green springs). Residues with similar counterparts in the kinases Erk2 and PKACA are given on a white background, while the grey background indicates dissimilar residues. The homology models of the enzymes *Tb*TryS (purple) abd *Tc*TryS (orange) were aligned with the MatchMaker methodology in UCSF Chimera.[6]

REFERENCES

[1] Benítez, D.; Medeiros, A.; Fiestas, L.; Panozzo-Zenere, E. A.; Maiwald, F.; Prousis, K. C.; Roussaki, M.; Calogeropoulou, T.; Detsi, A.; Jäger, T.; Šarlauskas, J.; Peterlin Mašič, L.; Kunick, C.; Labadie, G. R.; Flohé, L.; Comini, M. A. Identification of novel chemical scaffolds inhibiting trypanothione synthetase from pathogenic Trypanosomatids. *PLoS Neglected Trop. Dis.* **2016**, *10* (4), e0004617.

[2] Orban, O. C. F.; Korn, R. S.; Benítez, D.; Medeiros, A.; Preu, L.; Loaëc, N.; Meijer, L.; Koch, O.; Comini, M. A.; Kunick, C. 5-Substituted 3-chlorokenpaullone derivatives are potent inhibitors of *Trypanosoma brucei* bloodstream forms. *Bioorg. Med. Chem.* **2016**, *24* (16), 3790–3800.

[3] Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298* (5600), 1912–1934.

[4] Batista, J.; Hawkins, P. C. D.; Tolbert, R.; Geballe, M. T. SiteHopper - a unique tool for binding site comparison. *J. Cheminf.* **2014**, *6* (Suppl 1), P57.

[5] Eid, S.; Turk, S.; Volkamer, A.; Rippmann, F.; Fulle, S. KinMap: A web-based tool for interactive navigation through human kinome data. *BMC Bioinf.* **2017**, *18* (1), 16.

[6] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.

ACKNOWLEDGEMENTS

Finally, I would like to thank all the people who supported me throughout the projects of my thesis.

First, I thank Prof. Dr. Daniel Rauh for providing me with the opportunity to pursue my research goals under his supervision. I acknowledge the help of Prof. Dr. Stefan Kast and his readiness to review this thesis as the second referee. I am especially grateful to both of them and Dr. Schürmann for their support during the final months of writing.

Second, I thank Dr. Oliver Koch who admitted me to his group and provided interesting research topics.

I acknowledge the support by the Kekulé fellowship of the German FCI (Fonds der Chemischen Industrie).

I also thank Prof. Dr. Petra Mutzel for her trust and support during the development of the secondary structure assignment method SCOT in cooperation with Tobias Brinkjost.

Special thanks go to my colleague Tobias Brinkjost in good and bad times alike. I acknowledge his patience and fine grasp during the development and optimisation cycles of the two methods which were applied herein. Additionally, I acknowledge his support and comprehensive explanations concerning computer science, as well as the great lively debates on more general topics. In this context, I will probably never forget the perfect working atmosphere in our group which was granted by all the other wonderful, helpful, and cheering colleagues Lina Humbeck, Dr. Jette Pretzel, Anna Rudo, Dr. Hitesh Patel, Dr. Mauro Nogueira, Julia Jasper, and Sarah Albert. Many fruitful discussions and evening events often made my day. The former students and guests in our group often contributed to this good mood. Here, I particularly want to thank Nana Boateng for his first analyses of the TryS binding sites and Saskia Spitzer for her contribution in the application of the pocket prediction web server MetaPocket 2.0. I am also grateful to Laura Bewermeier who contributed to the preliminary analyses of binding site comparison methods. Finally, I thank Dr. Dennis M. Krüger for his help with the initial MD simulation studies.

Lina Humbeck and Tobias Brinkjost additionally supported this work by patiently and critically proof-reading parts of the manuscript of this work. Hats off to you.

Apart from that, I want to thank the researchers who provided us with their results and embraced the application of structure-based modelling approaches in the context of their scientific work: Dr. Lei Zhao and Prof. Dr. Yao-Wen Wu, Dr. Tom Mejuch and Prof. Dr. Herbert Waldmann, Dr. Helena Kaitsiotou and Prof. Dr. Daniel Rauh, Sara Maurer, Benjamin Buchmuller, and Prof. Dr. Daniel Summerer.

Coming to an end here, I realise that there are even more people at the TU Dortmund University to thank. I am grateful for all the smaller and larger contributions of the people who are too many to individually name them here.

I am especially grateful to my family and friends whose backing and encouragement was indispensable for this work. In particular, I want to thank my parents for their endless support and trust during the years of my studies.

Eidesstattliche Versicherung (Affidavit)

Ehrt, Christiane
Name, Vorname
(Surname, first name)

173378
Matrikel-Nr.
(Enrolment number)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden, § 63 Abs. 5 Hochschulgesetz NRW.

Die Abgabe einer falschen Versicherung an Eides statt ist strafbar.

Wer vorsätzlich eine falsche Versicherung an Eides statt abgibt, kann mit einer Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft werden, § 156 StGB. Die fahrlässige Abgabe einer falschen Versicherung an Eides statt kann mit einer Freiheitsstrafe bis zu einem Jahr oder Geldstrafe bestraft werden, § 161 StGB.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:

Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offences of this type is the chancellor of the TU Dortmund University. In the case of multiple or other serious attempts at deception, the candidate can also be unenrolled, Section 63, paragraph 5 of the Universities Act of North Rhine-Westphalia.

The submission of a false affidavit is punishable.

Any person who intentionally submits a false affidavit can be punished with a prison sentence of up to three years or a fine, Section 156 of the Criminal Code. The negligent submission of a false affidavit can be punished with a prison sentence of up to one year or a fine, Section 161 of the Criminal Code.

I have taken note of the above official notification.

Ort, Datum
(Place, date)

Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

Protein Binding Site Comparison -

The Impact of Binding Site Similarity on Hit Identification in Early Drug Discovery

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.
Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder der TU Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

I hereby swear that I have completed the present dissertation independently and without inadmissible external support. I have not used any sources or tools other than those indicated and have identified literal and analogous quotations.

The thesis in its current version or another version has not been presented to the TU Dortmund University or another university in connection with a state or academic examination.*

*Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the PhD thesis is the official and legally binding version.

Ort, Datum
(Place, date)

Unterschrift
(Signature)