Betrachtung der Ähnlichkeit von niedermolekularen Verbindungen unter Berücksichtigung der biologischen Aktivität

DISSERTATION

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Dr. rer. nat.)

der Fakultät für Chemie und Chemische Biologie
der Technischen Universität Dortmund

vorgelegt von
Dipl.-Pharm. Lina Humbeck
aus Marburg

1. Februar 2019

erster Gutachter: Prof. Dr. Daniel Rauh
zweiter Gutachter: Prof. Dr. Daniel Summerer

Die vorliegende Arbeit wurde in der Zeit von Februar 2014 bis Januar 2019 in der Arbeitsgruppe von Dr. Oliver Koch unter der Leitung von Prof. Dr. Daniel Rauh im Fachbereich Chemische Biologie der Fakultät Chemie und Chemische Biologie der Technischen Universität Dortmund angefertigt.

"To summarize: it is a well-known fact that those people who must want to rule people are, ipso facto, those least suited to do it. To summarize the summary: anyone who is capable of getting themselves made President should on no account be allowed to do the job. To summarize the summary of the summary: people are a problem."
Douglas Adams, The Restaurant at the End of the Universe

Inhaltsverzeichnis

Abkürzungsverzeichnis xiii

Publikationen xvii

1. Abstract and Outlook 1

2. Zusammenfassung und Ausblick 3

I. Einleitung und Aufgabenstellung 5
I.1. Ähnlichkeit von Molekülen 8
I.2. Entwickelte Konzepte 9
I.2.1. Der Maximum Similar Subgraph Ansatz (MSS-Ansatz) 11
I.2.2. Das Advanced Privileged Scaffold Concept (APSC) 11
I.3. Aufgabenstellung 12

II. Grundlagen 13
II.1. Repräsentation von Molekülen und Ähnlichkeitsfunktionen 15
II.1.1. Fingerprints 15
II.1.2. Exkurs: Graphtheorie 18
II.1.3. Repräsentation von Molekülen als Graphen - Molekülgraphen 21
II.1.4. Ähnlichkeitsfunktionen 24
II.2. Bromodomain-containing protein 4 (BRD4) und Peroxisome proliferator-activated receptor γ (PPARγ) 25
II.2.1. Bromodomain-containing protein 4 (BRD4) 25
II.2.2. Peroxisome proliferator-activated receptor γ (PPARγ) 27
II.2.3. Crosstalk von BRD4 und PPARγ 29
II.3. Benchmarkdatensätze 30
II.3.1. Benchmarkdatensatz von Stahl *et al.* (SBDS) 31

II.3.2. Benchmarkdatensatz von O'Boyle und Sayle (OBDS) 31
II.3.3. Benchmarkdatensatz von Backmann *et al.* (BBDS) 32

III. Ergebnisse und Diskussion 35

III.1. Erstellung und Analyse von CH*I*PMUNK 39
III.1.1. Einleitung 39
III.1.2. Ergebnisse und Diskussion 40
III.1.2.1. Physikochemische Eigenschaften von CH*I*PMUNK 42
III.1.2.2. Analyse des chemischen Raums von CH*I*PMUNK und Vergleich zum bioaktiven Raum 43
III.1.2.3. Vergleich von CH*I*PMUNK und DrugBank sowie iPPI 50
III.1.2.4. Anwendungsbeispiel 53
III.1.3. Zusammenfassung 54

III.2. Die Atomtypen- und Pharmakophor-Ebene der Ähnlichkeit 57
III.2.1. Einleitung 57
III.2.2. Ergebnisse und Diskussion 59
III.2.2.1. fconv-Atomtypen und fconv-Atomtypengruppen 59
III.2.2.2. Pharmakophor-Ebene 70
III.2.2.3. Analyse verschiedener Ähnlichkeitsfunktionen 73
III.2.2.4. Zulassung von Lücken 79
III.2.2.5. Clusteringergebnisse von StruClus unter Verwendung von verschiedenen Repräsentationen und Ähnlichkeitsebenen für den SBDS 82
III.2.3. Zusammenfassung 85

III.3. Die Scaffold- und Protein-Ebene der Ähnlichkeit 87
III.3.1. Einleitung 87
III.3.1.1. Privilegierte Scaffolds 88
III.3.1.2. Privilegierte Strukturen in Proteinen und das APSC 89
III.3.2. Ergebnisse und Diskussion 89
III.3.2.1. SPARROW-Analyse der DrugBank 89
III.3.2.2. Anwendung des APSC auf das Chromon-Scaffold 90
III.3.2.3. Analyse des Chromon-Scaffolds als erweitertes privilegiertes Scaffold 92
III.3.2.4. Ergebnisse des initialen Screens 98
III.3.2.5. Biochemische Charakterisierung initialer Hits 99
III.3.2.6. Biochemische Charakterisierung einer Substanzserie zur Inhibition von EGFR-Varianten 104

III.3.2.7. Biochemische Charakterisierung von PPARγ-Modulatoren . . 110
III.3.2.8. Ergebnisse der orthogonalen Assays 113
III.3.2.9. Kokristallstrukturen von BRD4 114
III.3.2.10. Bedeutung der entdeckten Ähnlichkeit von BRD4 und PPARγ: Resistenzen und neue therapeutische Strategien 117
III.3.3. Zusammenfassung 120

III.4. Die Bioisoster- und Shape-Ebene der Ähnlichkeit 123
III.4.1. Einleitung 123
III.4.2. Ergebnisse und Diskussion 125
III.4.2.1. Bioisoster-Ebene 125
III.4.2.2. Shape-Ebene 130
III.4.2.3. Herausforderungen und deren mögliche Lösungen 133
III.4.3. Zusammenfassung 138

IV. Material und Methoden 139
IV.1. Verbrauchsmittel und Chemikalien 141
IV.1.1. Allgemeine Anmerkungen 141
IV.1.2. Verbrauchsmittel 141
IV.1.3. Chemikalien, Proteine und Bakterienstämme 142
IV.2. Geräte 142

IV.1. Labor Experimente 145
IV.1.1. Proteinexpression von BRD4 145
IV.1.2. Proteinaufreinigung von BRD4 145
IV.1.3. Proteinkristallographie von BRD4 146
IV.1.4. Röntgenstrukturanalyse 146
IV.1.5. Assaysysteme 147
IV.1.5.1. BRD4 homogeneous time-resolved Förster-Resonanzenergietransfer (HTRF) Assay 147
IV.1.5.2. BRD4 Microscale Thermophoresis (MST) Assay 149
IV.1.5.3. BRD4 Thermal Shift Assay (TSA) 150
IV.1.6. Synthese von 4-Amino-2-hydroxy-5-[4-(2-pyridylsulfamoyl)-phenyldiazenyl]-benzoesäure (SSLH01, **21**) 151

IV.2. *In silico* Experimente 153
IV.2.1. Erstellung von Plots und Histogrammen 153
IV.2.2. Erstellung von Proteinabbildungen 153
IV.2.3. Erstellung von Abbildungen von Reaktionen und niedermolekularen Verbindungen 153

IV.2.4. Erstellung von Workflows . . . 153
IV.2.5. Vorbereitung von Molekülen und Reaktionen zur Erstellung von CH*I*PMUNK . . . 154
IV.2.6. Verwendete Reaktionen zur Erstellung von CH*I*PMUNK . . . 154
IV.2.7. Berechnung der physikochemischen Eigenschaften von CH*I*PMUNK . 154
IV.2.8. Clustering . . . 155
IV.2.8.1. Qualitätskriterien . . . 155
IV.2.8.2. StruClus Vortest und Test verschiedener Molekülrepräsentationen auf dem SBDS . . . 157
IV.2.8.3. StruClus-Clustering von CH*I*PMUNK . . . 157
IV.2.8.4. Visualisierung der Clustering-Ergebnisse . . . 158
IV.2.9. KNIME-Workflow zur Suche nach Molekülen in CH*I*PMUNK . . . 158
IV.2.10. Vorbereitung der Moleküle für den BBPMCS-Algorithmus . . . 159
IV.2.11. Analyse des MSS-Ansatzes mithilfe des OBDS . . . 159
IV.2.11.1. Änlichkeitssuche mittels ECFP4 und Tanimoto-Koeffizient (TC) 160
IV.2.11.2. Änlichkeitssuche mittels BBPMCS-Algorithmus . . . 160
IV.2.12. Bindetaschenvergleich der APSC-Proteine . . . 161
IV.2.13. Docking . . . 161
IV.2.13.1. Vorbereitung der Moleküle für das Docking . . . 162
IV.2.13.2. Vorbereitung der Proteine für das Docking . . . 162
IV.2.14. Selektion von zu APSC Liganden ähnlichen Verbindungen . . . 162
IV.2.15. Suche nach Liganden von PPARγ oder BRD4 . . . 164
IV.2.16. Ermittlung der bezüglich EGFR-BRD4-Inhibition sensitiven und insensitven Zelllinien . . . 164
IV.2.17. Analyse und Vergleich von bezüglich EGFR-BRD4-Inhibition sensitiven und insensitven Zelllinien . . . 165
IV.2.18. Erstellung der Molekülfragmente für den Shape-Vergleich . . . 166
IV.2.19. *In silico* Erstellung der DNA-kodierten Bibliotheken . . . 166
IV.2.20. Shape-Vergleich . . . 167
IV.2.21. Erstellung der Bioisostere . . . 168

V. Literaturverzeichnis 169

Anhang iii

A. Die Atomtypen- und Pharmakophor-Ebene der Ähnlichkeit iii
A.1. Zuordnung der fconv-Atomtypengruppen zu den fconv-Atomtypen . . iii
A.2. Verteilung der Scores von ECFP4-TC-Ansätzen und MSS-Ansatz . . . xi
A.3. Korrelation der Scores von ECFP4-TC-Ansätzen und MSS-Ansatz . . xii

A.4. Korrelation des Spearman-Rangkorrelationskoeffizienten mit der Graphgröße . . . xiii
A.5. Abkürzungen der Fingerprints des OBDS . . . xiv
A.6. Vergleich der Ergebnisse der Proof of Concept Studie . . . xv
A.7. Analyse der diskriminativen Kraft an einem Beispiel aus der Proof of Concept Studie . . . xvii
A.8. Ergebnisse der MSS-Suchen für den OBDS . . . xviii
A.8.1. fconv-Atomtypen- und fconv-Atomtypengruppen-Ebene inklusive Test des ct-Parameters . . . xviii
A.8.2. fconv-Atomtypengruppen-Ebene inklusive Ähnlichkeitsmatrix 1 . . . xx
A.8.3. fconv-Atomtypengruppen-Ebene inklusive Ähnlichkeitsmatrix 2 . . . xxi
A.8.4. Pharmakophor-Ebene . . . xxii
A.8.5. Pharmakophor-Ebene inklusive Ähnlichkeitsmatrix 2 . . . xxiii
A.8.6. Analyse verschiedener Ähnlichkeitsfunktionen . . . xxiv
A.8.7. Zulassung von Lücken . . . xxviii
A.9. Clusteringergebnisse von StruClus unter Verwendung von verschiedenen Repräsentationen und Ähnlichkeitsebenen für den SBDS . . . xxx

B. Die Scaffold- und Protein-Ebene der Ähnlichkeit xxxiii
B.1. Alignment der Proteinstrukturen aus dem APSC . . . xxxiii
B.1.1. Globales Alignment mittels Clustal Omega . . . xxxiii
B.1.2. Phylogenetischer Baum . . . xxxvi
B.1.3. Ergebnisse eines lokalen Alignments mittels BLAST . . . xxxvii
B.2. Analyse des Chromon-Scaffolds als erweitertes privilegiertes Scaffold . . . xxxviii
B.3. Ergebnisse des initialen Screens . . . xliii
B.4. Biochemische Charakterisierung einer Substanzserie zur Inhibition von EGFR-Varianten . . . xliv
B.5. Analyse der zellulären Effekte der BRD4-Inhibitoren aus der EGFR-Inhibitor-Serie . . . l
B.6. Kokristallstrukturen von BRD4 . . . li
B.6.1. Datensammlungs- und Verfeinerungsstatistiken der BRD4(1)-Komplexstrukturen . . . li
B.7. BRD4(1)-Sequenzen und Eigenschaften . . . lii
B.7.1. BRD4(1)-Sequenz mit His-Tag . . . lii
B.7.2. BRD4(1)-Sequenz ohne His-Tag . . . liii
B.8. Zusammensetzungen von Puffern . . . liii
B.8.1. Lysis-Puffer . . . liii
B.8.2. Größenausschluss-Puffer . . . liii
B.8.3. TSA-Puffer . . . liii

B.8.4. Reservoirlösungen liii

Danksagungen **lv**

Eidesstattliche Versicherung **lvi**

Abbildungsverzeichnis

I.0.1. Veranschaulichung der Ähnlichkeit anhand von Beispielen 10
II.0.1. Fingerprintrepräsentationen von Molekülen 17
II.0.2. Graphrepräsentationen von Molekülen 22
II.0.3. Die BET-Familie 25
II.0.4. Funktion von BRD4 26
II.0.5. BETi in klinischen Studien 27
II.0.6. Funktionen von PPARγ 28
II.0.7. Isosilybin A 28
II.0.8. Aufbau des OBDS 32

III.1.1. Aufbau von CH*I*PMUNK 40
III.1.2. Eigenschaftshistogramme der CoMol-Subsets von CH*I*PMUNK 44
III.1.3. Eigenschaftshistogramme der CoMol_{500}-Subsets von CH*I*PMUNK . . . 45
III.1.4. Eigenschaftshistogramme der ZINC-Subsets von CH*I*PMUNK 46
III.1.5. Eigenschaftshistogramme der ZINC_{500}-Subsets von CH*I*PMUNK 47
III.1.6. Analyse des chemischen Raumes von CH*I*PMUNK 48
III.1.7. Analyse der Ähnlichkeit von CH*I*PMUNK zur DrugBank 51
III.1.8. Analyse der Ähnlichkeit von CH*I*PMUNK zur iPPI 52
III.1.9. Strukten von Citalopram und iPPI 463 53
III.1.10. Strukten von Alprazolam und Triazolam 54

III.2.1. Beispiel für zwei Moleküle mit gleichem Pharmakophor 58
III.2.2. Ausgewählte Hits einer MSS-Suche 60
III.2.3. Vergleich der Ränge der Proof of Concept Studie 62
III.2.4. Per ECFP4-TC nicht unterscheidbare Hits einer MSS-Suche 63
III.2.5. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Element-Ebene 66
III.2.6. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypen- und Atomtypengruppen-Ebene (Ausschnitt) 67
III.2.7. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypengruppen-Ebene und Ähnlichkeitsmatrix 1 (Ausschnitt) 68
III.2.8. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypengruppen-Ebene und Ähnlichkeitsmatrix 2 (Ausschnitt) 69

III.2.9. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Pharmakophor-Ebene (Ausschnitt) 70
III.2.10. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Pharmakophor-Ebene und Ähnlichkeitsmatrix 2 (Ausschnitt) 71
III.2.11. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 1 bis 4 (Ausschnitt) . . . 76
III.2.12. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 5 und 6 (Ausschnitt) . . 77
III.2.13. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 7 ud 8 (Ausschnitt) . . . 78
III.2.14. Beispiele für ähnliche Moleküle mit Variation der Kettenlänge 79
III.2.15. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS Single-Assay Benchmarkdatensatz unter Zulassung von Lücken (Ausschnitt) 80
III.2.16. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS Multi-Assay Benchmarkdatensatz unter Zulassung von Lücken (Ausschnitt) 81

III.3.1. Ergebnis von SPARROW: das erweiterte privilegierte Chromon-Scaffold 91
III.3.2. Bioaktive Moleküle mit Chromon-Scaffold 93
III.3.3. Strukturen von BRD4, PPARγ, CDK2, ESR2 sowie NR1H4 94
III.3.4. Ergebnis des Bindetaschenvergleichs mit IsoMIF 95
III.3.5. Alignment von BRD4 mit PPARγ, CDK2, ESR2 sowie NR1H4 mittels IsoMIF basierend auf vollständigen Liganden 97
III.3.6. Bekannter dualer EGFR-BRD4-Inhibitor 104
III.3.7. Strukturen von RL078, RL077 und RL046 106
III.3.8. Strukturen von RL129, RL130, RL070 und RL100 106
III.3.9. Strukturen von Sulfasalazin, MS436 und SSLH01 112
III.3.10. Komplexstrukturen von BRD4 mit den Verbidnungen **2**, **3** und **4** 115
III.3.11. Komplexstrukturen von BRD4 mit Sulfasalazin und SSLH01 116
III.3.12. Alignment der BRD4-Sulfasalazin-Komplexstruktur mit BRD4-MS436- und BRD4-biBET-Komplexstrukturen 116
III.3.13. Zusammenspiel von BRD4 und PPARγ 119

III.4.1. Beispiel für ein MMP 124
III.4.2. Zwei Beispiele für bioisostere Paare mit Morpholin-Rest 126
III.4.3. Zwei Beispiele für bioisostere Paare mit einem Kohlenstoffatom 127
III.4.4. Zwei Beispiele für bioisostere Paare mit einem Wasserstoffatom 128
III.4.5. Zwei Beispiele für bioisostere Paare mit zyklischen Gruppen 129
III.4.6. Beispiel für einen Shape-Vergleich mittels NPR 132
III.4.7. Beispiel für einen Substitutionspfad 133

III.4.8. Beispiel für überlappende austauschbare Substrukturen 133
III.4.9. Beispiel für eine Äquivalenzklasse mit drei Ringsystemen 134
III.4.10. Beispiel für eine Äquivalenzklasse mit zwei Ringsystemen 134
III.4.11. Beispiel für eine Äquivalenzklasse mit sechs Ringsystemen 134
III.4.12. Beispiele für Protomere 136
III.4.13. Beispiele für Tautomere 137

IV.1.1. HTRF-Assay Prinzip 148
IV.1.2. Struktur von SSLH01 (**21**) 151

IV.2.1. Beispiel eines KNIME-Workflows 159
IV.2.2. Das APSC und seine Validierung 163
IV.2.3. Ausgewählte DEL-kompatible Reaktionen 166

A.1. Score-Verteilung von ECFP4-TC-Ansätzen und MSS-Ansatz xi
A.2. Korrelation der Scores von ECFP4-TC-Ansätzen und MSS-Ansatz . . . xii
A.3. Korrelation des Spearman-Rangkorrelationskoeffizienten mit der Graphgröße xiii
A.4. Per ECFP4-TC nicht unterscheidbare Hits und relevante weitere Strukturen einer MSS-Suche xvii
A.5. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypen- und Atomtypengruppen-Ebene mit Zeitbegrenzung von 1000 Millisekunden xviii
A.6. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypen- und Atomtypengruppen-Ebene mit Zeitbegrenzung von 2000 Millisekunden xix
A.7. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypengruppen-Ebene und Ähnlichkeitsmatrix 1 . xx
A.8. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypengruppen-Ebene und Ähnlichkeitsmatrix 2 . xxi
A.9. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Pharmakophor-Ebene xxii
A.10. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Pharmakophor-Ebene und Ähnlichkeitsmatrix 2 . . . xxiii
A.11. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 1 und 2 xxiv
A.12. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 3 und 4 xxv
A.13. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 5 und 6 xxvi
A.14. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 7 ud 8 xxvii

A.15. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS Single-Assay Benchmarkdatensatz unter Zulassung von Lücken xxviii
A.16. Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS Multi-Assay Benchmarkdatensatz unter Zulassung von Lücken xxix
A.17. Clusteringergebnisse auf dem SBDS xxx

B.1. Phylogenetischer Baum xxxvi
B.2. Ligand-fokussierte Ansicht der Strukturen von BRD4 und ESR2 xxxviii
B.3. Aligment von BRD4 mit PPARγ, CDK2, ESR2 sowie NR1H4 mittels IsoMIF basierend auf dem Scaffold xxxix
B.4. Komplexstrukturen von BRD4 und PPARγ mit selektiven Liganden . . xl
B.5. Interaktionen des Chromon-Scaffolds mit BRDs xli
B.6. Interaktionen des Chromon-Scaffolds mit CDK2 und CDK9 xlii
B.7. Ähnlichkeitsanalyse von BRD4-Hits aus dem APSC und dem Docking-Ansatz sowie Vergleich zu bekannten BRD4- und PPARγ-Liganden . . xliii
B.8. Strukturen der Positivontrollen **24** und **25** xliv
B.9. Alignment der BRD4-Sulfasalazin-Komplexstruktur mit relevanten weiteren Komplexstrukturen lii

Tabellenverzeichnis

II.0.1. Zusammensetzung des SBDS 31

III.1.1. Eigenschaften von CH*I*PMUNK basierend auf CoMol-Edukten 42
III.1.2. Eigenschaften von CH*I*PMUNK basierend auf ZINC-Edukten 42

III.2.1. Ergebnisse der Proof of Concept Studie 64
III.2.2. Ähnlichkeitsmatrix 1 68
III.2.3. Ähnlichkeitsmatrix 2 69
III.2.4. Ähnlichkeitsfunktionen 75
III.2.5. Ergebnisse von StruClus auf dem SBDS (gewichtete Gesamt-Reinheit und F1-Score) 84

III.3.1. Ergebnisse des initialen BRD4-Screens 98
III.3.2. Ergebnisse ausgewählter initialer Hits 103
III.3.3. BRD4-Hits aus der EGFR-Inhibitor-Serie 105
III.3.4. Zelluläre Charakterisierung von BRD4-Inhibitoren aus der EGFR-Inhibitor-Serie 108
III.3.5. Ergebnisse der PPARγ-Modulatoren 111

A.1. Zuordnung der fconv-Atomtypengruppen zu den fconv-Atomtypen . . . x
A.2. Übersicht der Fingerprints des OBDS xiv
A.3. Erweiterte Ergebnisse der Proof of Concept Studie xvi
A.4. Ergebnisse von StruClus auf dem SBDS (Gesamt-Reinheit und Clusteranzahl) xxxi

B.1. BLAST Ergebnis für APSC-Proteine: maximaler und totaler Score . . . xxxvii
B.2. BLAST Ergebnis für APSC-Proteine: Suchsequenz-Abdeckung xxxvii
B.3. BLAST Ergebnis für APSC-Proteine: E-Wert xxxvii
B.4. BLAST Ergebnis für APSC-Proteine: Identität xxxvii
B.5. Ergebnisse der EGFR-Inhibitor-Serie für BRD4 xlix
B.6. Häufig mutierte Gene in gegenüber EGFR-BRD4-Inhibiton besonders sensitiven Zelllinien l
B.7. Datensammlungs- und Verfeinerungsstatistiken li

Abkürzungsverzeichnis

AChE Acetylcholinesterase 77, 120

ADMET *absorption, distribution, metabolism, excretion, toxicology* (Resorption, Verteilung, Metabolismus, Ausscheidung, Toxikologie) 117

APSC *advanced privileged scaffold concept* (erweiterte privilegierte Scaffold Konzept) xxxi, xxxv, xxxvi, xxxviii, xlii, 3, 4, 7, 8, 10, 81–84, 86, 88–90, 92–96, 107, 108, 111, 114, 115, 151–154

AUC *area under the curve* (Fläche unter der Kurve) 15

BBP *block and bridge preserving* (Block und Brücken beibehaltend) 18

BBPMCS *block and bridge preserving maximum common subgraph* (BBP größter gemeinsamer Teilgraph) xvii–xxviii, 19, 21, 29, 52, 53, 57, 59–65, 67–75, 123, 131, 149, 150

BET *bromodomain and extra terminal domain* 23

BETi Inhibitoren der BET-Proteinfamilie 24, 25, 101, 103, 111–113

BRD4 *bromodomain-containing protein 4* (Bromodomän-enthaltendes Protein 4) xxxi, xxxvii–xxxix, xlii, xlviii–li, 4, 23–25, 27, 28, 86, 88–93, 98, 99, 101–108, 110–115, 136–141, 151, 153–155

bRo5 *beyond rule of five* (außerhalb der Lipinski-Regel) 36–38, 45, 49

CDK2 *cyclin-dependent kinase 2* (Cyklin-abhängige Kinase 2) xxxvii, xxxviii, xli, 77, 84, 86, 88, 90–93, 111, 113, 114, 151, 152

COMAS *Compound Management and Screening Center* 92, 114, 152, 153

COX-1 Cyclooxygenase-1 121

COX-2 Cyclooxygenase-2 29, 77, 121

CSD *Cambridge Structural Database* 123, 124, 127, 131, 156

DC Dice-Koeffizient 22

DEL *DNA encoded library* (DNA-kodierte Bibliothek) 124, 125, 131, 156

DMSO Dimethylsulfoxid li, 108, 134, 139, 141

E. coli *Escherichia coli* 136

ECFP4 *extended-connectivity fingerprint 4* x, xi, xv, xvi, xlii, 14, 15, 44–47, 53–58, 67, 79, 93, 150, 152

EGFR *epidermal growth factor receptor* (Epidermaler Wachstumsfaktorrezeptor) xliii, xlviii, xlix, 98, 99, 101–103, 111–114, 119, 154, 155

eRo5 *extended rule of five* (erweiterte Lipinski-Regel) 38

ESR2 Estrogenrezeptor β xxxvii, xxxviii, 84, 86, 88, 90–93, 114, 151, 152

fconv *file conversion* ii, ix–xi, xv, xvii, xix, xx, 51, 53, 54, 56, 57, 59, 61, 64, 67, 68, 76, 77, 127

FRET Förster-Resonanzenergietransfer 138

GAFF *general AMBER force field* 51

GST Glutathion-S-Transferase 139

HEPES 2-(4-(2-Hydroxyethyl)-1-piperazinyl)-ethansulfonsäure lii

HPLC *high performance liquid chromatography* (Hochleistungsflüssigkeitschromatographie) 133, 135, 142

HTRF *homogeneous time-resolved* Förster-Resonanzenergietransfer 4, 92, 93, 97–99, 105, 107, 108, 138, 139, 153

IPTG Isopropyl-β-D-thiogalactopyranosid 136

IUPAC *International Union of Pure and Applied Chemistry* 51

LB-Medium Luria-Bertani Medium 136

MCS *maximum common subgraph* (größter gemeinsamer Teilgraph) 18, 19, 69

MMP *matched molecular pair* 116–118, 158

MSS *maximum similar subgraph* (größter ähnlicher Teilgraph) x–xii, xv–xvii, 3, 4, 7, 8, 10, 19, 23, 28, 53–58, 61, 66, 67, 73, 77, 79, 80, 149

MST *Microscale Thermophoresis* 4, 82, 97, 105–107, 114, 135, 138, 140, 141

NFκB *nuclear factor 'kappa-light-chain-enhancer' of activated B-cells* 27

NPR *normalized principal moments of inertia ratio* 117, 123–125, 131, 157

NR1H4 *bile acid receptor* (Gallensäurerezeptor) xxxvii, xxxviii, 86, 88–93, 114, 151, 152

PAINS *pan-assay interference structures* 84, 156

PCA *principal component analysis* (Hauptkomponentenanalyse) 124

PMI *principal moments of inertia* (Hauptträgheitsmomente) 157

PPARγ *peroxisome proliferator-activated receptor γ* (Peroxisom-Proliferator-aktivierter Rezeptor γ) xxxvii–xxxix, xlii, 4, 23, 25–28, 86, 88–97, 104–107, 111–115, 151, 152, 154

PPII Protein-Protein-Interaktions-Inhibitor 36, 38, 45, 48, 49

PROTAC *proteolysis targeting chimera* (Proteolyse-vermittelnde Chimäre) 8, 112

PSA *polar surface area* (polare Oberfläche) 36, 37, 144

RMSD *root mean square deviation* (Wurzel der mittleren Abweichungs-Quadrate) 109, 110

Ro5 *rule of five compliant* (innerhalb der Lipinski-Regel) 35–38

ROC *Receiver-Operating-Characteristic* 15

SA *synthetic accessibility* (synthetische Zugänglichkeit) 35–37, 49, 144, 145

SPP *similar property principle* 6, 7, 28

TC Jaccard-Tanimoto-Koeffizient (kurz Tanimoto-Koeffizient) x, xi, xv, xvi, xlii, 22, 23, 45–48, 53–58, 67, 79, 93, 150, 152

TCEP Tris(2-carboxyethyl)phosphin lii

TSA *Thermal Shift Assay* lii, 4, 82, 97, 106, 107, 114, 138, 141

VEGFR1 *vascular endothelial growth factor receptor 1* (Vaskulärer endothelialer Wachstumsfaktor-Rezeptor 1) 119

Publikationen

Im Rahmen dieser Arbeit wurden folgende Publikationen veröffentlicht:

- **Lina Humbeck** und Oliver Koch: What can we learn from bioactivity data? Chemoinformatics tools and applications in chemical biology research, *ACS Chemical Biology*, 12: 23-35; **2017**
- Till Schäfer*, Nils Kriege*, **Lina Humbeck***, Karsten Klein, Oliver Koch, Petra Mutzel: Scaffold Hunter: A comprehensive visual analytics framework for drug discovery, *Journal of Cheminformatics*, 9: 28; **2017**
- Nils Kriege, **Lina Humbeck**, Oliver Koch: Chemical similarity and substructure searches, *in Ranganathan, S., Nakai, K., Schönbach C. und Gribskov, M. (eds.) Encyclopedia of Bioinformatics and Computational Biology, Volume 2, Seiten 640-649*; **2019**; Oxford: Elsevier
- **Lina Humbeck**, Sebastian Weigang, Till Schäfer, Petra Mutzel, Oliver Koch: CH*I*PMUNK: A virtual synthesizable small-molecule library for medicinal chemistry, exploitable for protein–protein interaction modulators, *ChemMedChem*, 13: 532-539; **2018**
- Julia B. Jasper, **Lina Humbeck**, Tobias Brinkjost, Oliver Koch: A novel interaction fingerprint derived from per atom score contributions: Exhaustive evaluation of interaction fingerprint performance in docking based virtual screening, *Journal of Cheminformatics*, 10: 15; **2018**
- **Lina Humbeck**, Jette Pretzel, Saskia Spitzer, Oliver Koch: Discovery of an unexpected similarity in ligand binding between BRD4 and PPARγ (submitted)

* Diese Autoren haben gleichermaßen zur Publikation beigetragen.

1. Abstract and Outlook

The acceleration of drug development and the improvement of therapy options are enormous challenges with increasing demand for interdisciplinary solutions. One already very successfully applied strategy is the rational computer-aided design of drug candidates in combination with other, e.g., biochemical, analyses. Nevertheless, the potential is still unexhausted and there is a huge need for overlapping research in the areas of computer science, medicine, pharmacy, biology and chemistry. One fundamental but rather underexplored basis for rational design of innovative drugs is the suitable recognition of the similarity of small molecules, particularly of the biosimilarity. Maggiora et al. (2014) formulated as a pioneering challenge in the field of similarity analyses of small molecules the development of concepts that consider biosimilarity. The work in hand meets this challenge and investigates the topic of similarity of small molecules with respect to their bioactivity on different levels utilizing a repertoire of chemoinformatics, biochemical and structural biology methods. Thereby, various insights into the complex field of biosimilarity could be achieved. Two methods for the recognition and assessment of similarity were developed, the maximum similar subgraph approach (MSS approach) and the advanced privileged scaffold concept (APSC).

Different levels of similarity (atom type groups, pharmacophores, bioisosteres and shape) were analyzed for the MSS approach. In doing so, general rules were established which preserve the biosimilarity despite structural changes, the MSS approach was evaluated and suitable default parameters for the similarity search with the search algorithm (BBPMCS) of the MSS approach were identified. Additionally, the competitiveness of the MSS approach with respect to established methods was demonstrated in a large-scale benchmark study. It could be shown that the biosimilarity decisively depends on the target. Accordingly, the newly developed MSS approach enables to optimize the assessment of biosimilarity for each protein. Moreover, the limitations of the current realization were recognized and possibilities for further improvement, e.g., the integration of the three-dimensional structure and related aspects were discussed.

As a second method, the APSC which is based on the scaffold level of similarity could be successfully introduced and validated by means of bromodomain-containing protein 4 (BRD4), a recent drug target, which regulates the transcription of genes. The APSC-based analysis led to the identification of novel BRD4 inhibitors which were characterized biochemically by multiple assay systems (homogeneous time-resolved Förster resonance energy transfer (HTRF), Microscale Thermophoresis (MST) and Thermal Shift Assay (TSA)). The cocrystallization of five BRD4 inhibitors was successful.

Furthermore, a similarity of BRD4 and peroxisome proliferator-activated receptor γ (PPARγ), a nuclear receptor and transcription factor, was discovered. The similarity of these two drug targets can be utilized for the design of dual modulators that have the potential to on the one hand be effective on resistant tumor cells (e.g., BETi resistant tumor cells) and on the other hand prevent resistance development. Through this analysis, not only small molecule similarity (ligand similarity) and big molecule similarity (protein similarity) but also computer-based and wet lab experiments were bridged.

Furthermore, a database (CH*I*PMUNK) was created which serves as a benchmark dataset for the assessment of new algorithms and as an idea generator for drug design. CH*I*PMUNK features a broad chemical space and was successfully partitioned into chemically and biologically meaningful clusters by a clustering algorithm (StruClus).

In this interdisciplinary work it could be demonstrated that it is possible, e.g., through the MSS approach or the APSC to assess molecules regarding their similarity from a biological perspective. This is an important foundation for the improvement of drug development because thereby the relevant factor biosimilarity can be integrated into the development process early on and a better, i.e., with bioactive molecules enriched, preselection of molecules to be analyzed is achieved. Finally, this saves resources like time and material and enhances the chances of successful drug development. The biosimilarity was evaluated primarily focussing on improving drug development. For this, the presented results including the MSS approach and the APSC can be used manifoldly, e.g., to identify new ligand series and to predict potential side effects, toxicological effects or the environmental impact.

As outlined in the present work, algorithms which process both, a representation of the small molecule as well as a representation of the respective target, are needed in future. The gained knowledge may result in in-depth insights into the relationship between the small molecule, protein structures and the bioactivity. To a limited extent, this is already possible by so-called proteochemometric approaches.

2. Zusammenfassung und Ausblick

Die Beschleunigung der Wirkstoffentwicklung und die Verbesserung der Therapiemöglichkeiten sind große Herausforderungen, die in zunehmenden Maß interdisziplinäre Lösungsstrategien erfordern. Eine bereits sehr erfolgreich angewandte Strategie ist das rationale, computer-gestützte Design von Wirkstoffkandidaten in Kombination mit weiteren, z. B. biochemischen, Analysen. Nichtsdestotrotz ist das Potenzial bei weitem noch nicht ausgeschöpft und es besteht ein großer Bedarf an überlappender Forschung in den Bereichen Informatik, Medizin, Pharmazie, Biologie und Chemie. Eine fundamentale, jedoch noch unzureichend umgesetzte, Grundlage für das rationale Design innovativer Wirkstoffe ist die adäquate Erfassung der Ähnlichkeit von niedermolekularen Verbindungen, insbesondere der Bioähnlichkeit. Maggiora *et al.* (2014) formulierten als zukunftsweisende Herausforderung in dem Feld der Ähnlichkeitsanalyse von niedermolekularen Verbindungen die Entwicklung von Konzepten zur Berücksichtigung von Bioähnlichkeit. Die vorliegende Arbeit stellt sich dieser Herausforderung und untersucht das Thema der Ähnlichkeit von niedermolekularen Verbindungen in Bezug auf ihre Bioaktivität auf verschiedenen Ebenen und interdisziplinär mithilfe eines Repertoires an chemieinformatischen, biochemischen und strukturbiologischen Methoden. Dadurch konnten vielfältige Einblicke in das komplexe Gebiet der Bioähnlichkeit gewonnen werden. Es wurden zwei Methoden zur Erfassung und Bewertung der Ähnlichkeit entwickelt, der *maximum similar subgraph* (größter ähnlicher Teilgraph) (MSS) Ansatz und das *advanced privileged scaffold concept* (erweiterte privilegierte Scaffold Konzept) (APSC).

Für den MSS-Ansatz wurden verschiedene Ebenen der Ähnlichkeit (Atomtypengruppen, Pharmakophore, Bioisostere und Shape (Form)) analysiert. Dabei wurden allgemeine Regeln etabliert, welche die Bioähnlichkeit bei Veränderung der Struktur erhalten sollen. Weiterhin wurde der MSS-Ansatz evaluiert und es wurden geeignete Startparameter für die Ähnlichkeitssuche mittels Suchalgorithmus (BBPMCS) im MSS-Ansatz identifiziert. Zudem konnte in einer großangelegten Benchmarkstudie die Konkurrenzfähigkeit des MSS-Ansatzes gegenüber bereits etablierten Methoden demonstriert werden. Es konnte gezeigt werden, dass die Bioähnlichkeit einer niedermolekularen Verbindung maßgeblich durch die Zielstruktur determiniert wird. Der neu entwickelte MSS-Ansatz ermöglicht dementsprechend, die Erfassung der Bioähnlichkeit für jedes Protein zu optimieren und anzupassen. Zugleich wurden allerdings ebenfalls noch Limitationen in der Umsetzung erkannt und Möglichkeiten zur weiteren Optimierung, wie z. B. die Einbeziehung der dreidimensionalen Struktur und den damit verbundenen Aspekten, diskutiert.

Als zweite Methode konnte das APSC, welches auf der Scaffold-Ebene der Ähnlichkeit aufgebaut wurde, erfolgreich anhand von *bromodomain-containing protein 4* (Bromodomän-enthaltendes Protein 4) (BRD4), einer aktuellen Arzneistoffzielstruktur zur Regulierung der Transkription von Genen, eingeführt und validiert werden. Die APSC basierte Analyse führte zur Identifizierung neuer BRD4-Inhibitoren, welche in mehreren Assaysystemen (*homogeneous time-resolved* Förster-Resonanzenergietransfer (HTRF), *Microscale Thermophoresis* (MST) und *Thermal Shift Assay* (TSA)) biochemisch charakterisiert wurden. Es gelang die Kokristallisation von fünf BRD4-Inhibitoren. Zudem wurde durch das APSC eine Ähnlichkeit zwischen BRD4 und *peroxisome proliferator-activated receptor* γ (Peroxisom-Proliferator-aktivierter Rezeptor γ) (PPARγ), einem nukleären Rezeptor und Transkriptionsfaktor, entdeckt. Die Ähnlichkeit der beiden Arzneistoffzielstrukturen kann für das Design von dualen Modulatoren ausgenutzt werden, welche das Potenzial besitzen, einerseits wirksam gegenüber resistenten Tumorzellen (z. B. BETi resistente Tumorzellen) zu sein und andererseits einer Resistenzentwicklung vorzubeugen. Mit dieser Analyse wurde nicht nur eine Brücke zwischen der Ähnlichkeit niedermolekularer Verbindungen (Ligandähnlichkeit) und der Ähnlichkeit großer Moleküle (Proteinähnlichkeit) geschlagen, sondern ebenfalls zwischen computer-basierten und Labor Experimenten.

Des Weiteren wurde eine Datenbank (CH*I*PMUNK) erstellt, welche als Benchmarkdatensatz zur Testung neuer Algorithmen und als Ideengenerator für das Wirkstoffdesign zur Verfügung steht. CH*I*PMUNK weist einen großen chemischen Raum auf und konnte erfolgreich mithilfe eines Clustering-Algorithmus (StruClus) in chemisch-biologisch sinnvolle Cluster eingeteilt werden.

In dieser interdisziplinären Arbeit konnte gezeigt werden, dass es z. B. mit dem MSS-Ansatz oder dem APSC möglich ist, Moleküle bezüglich ihrer Ähnlichkeit aus biologischer Sicht zu bewerten. Dies ist ein wichtiger Grundstein zur Verbesserung der Wirkstoffentwicklung, da somit bereits früh der relevante Faktor Bioähnlichkeit in den Entwicklungsprozess einfließen kann und eine bessere, d. h. an bioaktiven Molekülen angereicherte, Vorauswahl an zu untersuchenden Molekülen getroffen werden kann. Dies spart wiederum Ressourcen wie Zeit und Material ein und erhöht die Chancen auf eine erfolgreiche Wirkstoffentwicklung. Die Bioähnlichkeit wurde mit dem primären Fokus auf der Verbesserung der Arzneistoffentwicklung betrachtet. Dabei können die vorliegenden Ergebnisse inklusive der MSS-Methode sowie des APSC vielfältig genutzt werden z. B. zur Identifizierung neuer Ligandenserien und zur Vorhersage von potentiellen Nebenwirkungen, toxikologischen Effekten oder möglicher Einflüsse auf die Umwelt.

Für die Zukunft wäre es interessant, wenn Algorithmen neben der Repräsentation der niedermolekularen Verbindung ebenfalls eine Repräsentation der jeweiligen Zielstruktur als Eingabe erwarten würden. Die Erkenntnisse daraus könnten einen tieferen Einblick in die Beziehung zwischen Struktur der niedermolekularen Verbindung, Proteinstrukturen und der Bioaktivität gewähren. In einem noch begrenzten Maß ist dies bereits mittels des sogenannten Proteochemometrik-Ansatzes möglich.

Teil I.

Einleitung und Aufgabenstellung

Die Gesundheit ist für jeden einzelnen Menschen, und damit im erweiterten Sinne auch für das Funktionieren der menschlichen Gesellschaft, von zentraler Bedeutung. Sie ist Voraussetzung für viele weitere wichtige Faktoren wie z. B. wirtschaftlichen Erfolg. Die Erhaltung bzw. Wiedererlangung der Gesundheit nach Krankheit stellt daher neben der Wasserversorgung, der Ernährung und der Energieversorgung eine der größten aktuellen Herausforderung einer stetig wachsenden Weltbevölkerung dar. Zudem steigt die Lebenserwartung in vielen Regionen. Leider gibt es für eine Reihe von Krankheiten, zu denen unter anderem auch Krebs, Aids oder Ebola gehören, bis heute keine ausreichenden, adäquaten Behandlungsmöglichkeiten. Des Weiteren können Therapieansätze z. B. durch Resistenzbildung unwirksam werden. Dem steigenden, spürbar hohen Bedarf an innovativen und bezahlbaren Therapiemöglichkeiten steht allerdings eine Zunahme der Forschungs-, Entwicklungs- und Studien-Kosten pro neu zugelassenem Arzneistoff gegenüber (Scannell *et al.* 2012).
Eine Möglichkeit der Kostenreduktion besteht in der Anwendung von computergestützter, rationaler Wirkstoffforschung. Dabei handelt es sich um einen Vorfilterungsschritt, der die Menge an durch teure (Materialkosten und personeller Zeitaufwand) Labormethoden zu testenden potentiellen Wirkstoffvorläufern (Hits) eingrenzt und somit in der eingegrenzten Menge die Chance einen Hit zu finden erhöht. Eine essentielle Voraussetzung für die rationale Wirkstoffforschung ist, dass eine Beziehung zwischen der Struktur einer niedermolekularen Verbindung und deren Eigenschaften besteht, wie sie bereits 1868 von Brown und Fraser postuliert wurde (Brown und Fraser 1868). 1990 wurde ein weiteres zentrales Prinzip der rationalen Wirkstoffforschung, das *similar property principle* (SPP), von Johnson und Maggiora (Maggiora *et al.* 2014) formuliert. Das SPP bezieht den Aspekt der Ähnlichkeit ein und besagt, dass molekular ähnliche Moleküle mit höherer Wahrscheinlichkeit ähnliche Eigenschaften besitzen als molekular unähnliche Moleküle (Maggiora *et al.* 2014; Wilkins und Randić 1980; Martin, Kofron und Traphagen 2002). Gemeint ist mit den ähnlichen Eigenschaften zumeist die Fähigkeit die native Funktion eines Zielmoleküls (in der Regel ein Protein) zu beeinflussen. Dieses Prinzip ermöglicht die bekannten Eigenschaften von im Labor getesteten Verbindungen auf noch ungetestete, jedoch ähnliche Verbindungen zu extrapolieren. Es ist ein so elementares Prinzip, dass es nicht nur für die Entwicklung von Humanarzneimitteln sondern auch für die Entwicklung von Tierarzneimittel und Phytoeffektoren angewendet werden kann. Folglich ist es für die rationale Wirkstoffforschung essentiell, sich umfassend mit der Ähnlichkeit von Molekülen zu beschäftigen. Die vorliegende Arbeit befasst sich aus diesem Grund mit der Ähnlichkeit von niedermolekularen Verbindungen, wobei der Fokus auf der Bioähnlichkeit, d. h. ähnlicher biologischer Aktivität, liegt. Auf die Thematik der Ähnlichkeit wird im nächsten Abschnitt detaillierter eingegangen.

I.1. Ähnlichkeit von Molekülen

Ähnlichkeit ist ein abstraktes und breit angewendetes Konzept, welches sowohl von der Relation als auch von der zu betrachtenden Eigenschaft abhängt. Im Kontext der Arbeit ist die Relation gegeben durch den Vergleich mit einem Referenzmolekül. Wie so vieles liegt auch die Ähnlichkeit im Auge des Betrachters (Herndon und Bertz 1987) und ist somit subjektiv und schwer zu quantifizieren. Folglich gibt es keine allgemeine, generell gültige Ähnlichkeit, sondern sehr viele Arten, welche jeweils einen anderen Aspekt der Ähnlichkeit beleuchten. In der Chemieinformatik und der medizinischen Chemie wird häufig die molekulare Ähnlichkeit verwendet (Maggiora *et al.* 2014). Allerdings ist das Ziel nicht die molekulare Ähnlichkeit, sondern die biologische Ähnlichkeit, d. h. ähnliche biologische Aktivität (Bioähnlichkeit), der Moleküle (mithilfe von molekularer Ähnlichkeit nach dem SPP) zu erfassen. Diese Diskrepanz zwischen Surrogatparameter (molekulare Ähnlichkeit) und Phänomen (Bioähnlichkeit) stellt eine bedeutende Herausforderung der Chemieinformatik und medizinischen Chemie dar. 2014 formulierten Maggiora *et al.* (2014) diese Herausforderung: „The development of computational similarity methods with reduced compound class dependence will be an important topic for future research. In addition, the exploration of new concepts to account for biological similarity of small compounds will be equally attractive.“ Diese Herausforderung ist das zentrale Thema der vorliegenden Arbeit, wobei auf die Einbeziehung der Bioähnlichkeit in die molekulare Ähnlichkeit fokussiert wird. Dazu erfolgte die Entwicklung und Analyse zweier neuer computer-basierter Konzepte zur Berücksichtigung der Bioähnlichkeit in der Betrachtung der Ähnlichkeit von niedermolekularen Verbindungen. Die beiden Konzepte, *maximum similar subgraph* Ansatz (MSS-Ansatz) und *advanced privileged scaffold concept* (APSC) werden unter Abschnitt I.2 genauer vorgestellt.

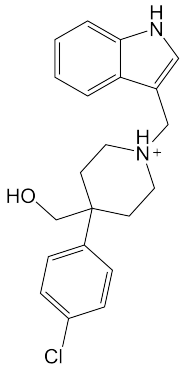
Die Betrachtung der Bioähnlichkeit muss logischerweise die Wirkung auf die Funktionalität des Zielproteins einbeziehen (die Ähnlichkeit liegt wie bereits zitiert im Auge des Betrachters, also des die niedermolekulare Verbindung bindenden Zielproteins). Deswegen werden in dieser Arbeit zur Erfassung der Bioähnlichkeit in Erweiterung früherer Konzepte beide Seiten, die molekulare Ähnlichkeit und die daraus resultierenden ähnlichen funktionellen Effekte auf die Zielproteine, im Zusammenhang betrachtet.

Dazu sollen folgende Aspekte der Bioähnlichkeit betrachtet werden: Atomtypengruppen, pharmakophore Eigenschaften, Scaffolds (Grundgerüste), Bioisostere und Shape (Form). Zwei Moleküle sind bioähnlich, wenn sie ähnliche biologische Effekte hervorrufen. Oft hängt dabei die Bioähnlichkeit eng mit der Ähnlichkeit bezüglich physikochemischer und molekularer Eigenschaften zusammen. Jedoch bedeutet eine Ähnlichkeit in physikochemischen und/oder molekularen Eigenschaften nicht zwangsläufig eine Bioähnlichkeit. Führen beispielsweise bereits kleine Veränderungen der Molekülstruktur zu einer drastischen Änderung des biologischen Effekts, so wird von *activity cliffs* gesprochen. Ein Beispiel für einen *activity cliff* sind die in den Abbildungen I.0.1a und I.0.1b dargestellten Dopamin D3-Rezeptor Antagonisten.

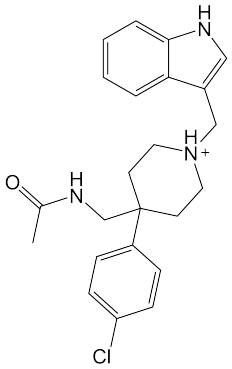
Ein weiterer zu beachtender Punkt ist, dass Moleküle ihre Wirkung auf verschiedene Weisen erzielen können, z. B. durch unterschiedliche Interaktionen mit dem Zielprotein (kompetitiv, kovalent, kovalent-reversibel, Bindung eines *proteolysis targeting chimera* (Proteolyse-vermittelnde Chimäre) (PROTAC)) und Interaktionen mit räumlich getrennten Bereichen (Substrat-, Kofaktor-Bindestelle, allosterische Bindestelle, katalytisches Zentrum). Auch Moleküle mit distinkter Struktur, wie z. B. die Estrogrenrezeptor α Agonisten Estradiol und Diethylstilbestrol (Abbildungen I.0.1c und I.0.1d) können bioänlich sein. Zudem kann bei gleichzeitig vorliegender molekularer Ähnlichkeit und Bioähnlichkeit das räumliche Alignment der ähnlichen Substrukturen des Moleküls deutlich verschieden ausfallen, wie das Beipiel von Dihydrofolsäure und Methotrexat (Abbildungen I.0.1e, I.0.1f und I.0.1g) zeigt. Zur Erfassung der Bioähnlichkeit ist es unumgänglich, den Ähnlichkeitsbegriff im medizinisch-chemischen Kontext zu erweitern und neue Algorithmen zur Erfassung und Diskriminierung dieser erweiterten Ähnlichkeit zu entwickeln. In der vorliegenden Dissertation wird dafür eine wichtige Pionierarbeit geleistet.

I.2. Entwickelte Konzepte

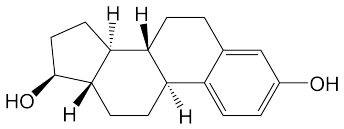
Zur computer-gestützten Analyse der Ähnlichkeit ist eine sowohl für den Computer geeignete als auch (bio)chemisch relevante Repräsentation des Moleküls notwendig. Dazu werden häufig Fingerprints verwendet (siehe Abschnitt II.1.1), da sie in der Regel schnell kalkulierbar sind und anschließende Analysen, wie Ähnlichkeitssuchen, ebenfalls zeit-effizient durchgeführt werden können. Allerdings ist die Umwandlung eines Moleküls in einen Fingerprint häufig mit einem Informationsverlust einhergehend. So wird oft die Topologie des Moleküls nur unvollständig im Fingerprint abgebildet. Dieser Informationsverlust wirkt sich, insbesondere bei der Bewertung von einer großen Anzahl von Molekülen (Big Data), durch eine geringere diskriminative Kraft negativ auf Ähnlichkeitsanalysen aus. Um den Informationsverlust zu minimieren, wird in dieser Arbeit eine Graph-basierte Repräsentation der Moleküle verwendet, deren Nutzen in einer Proof of Concept Studie unter Abschnitt III.2.2.1 gezeigt wird. Es wurden zwei Methoden entwickelt, welche in den nachfolgenden Abschnitten näher beschrieben werden. Die erste Methode, der MSS-Ansatz, integriert verschiedene Eben der Bioähnlichkeit in die Graph-basierten Ähnlichkeitsalgorithmen. Die zweite Methode, das APSC, integriert in die Ähnlichkeitsbetrachtung der niedermolekularen Verbindungen die Ähnlichkeit der Zielstrukturen. Während der MSS-Ansatz im Extremfall nur eine aktive Verbindung und eine Datenbank mit zu analysierenden Verbindungen benötigt, sind für das APSC umfassendere Daten, wie Liganden von weiteren Proteinen erforderlich. Damit lassen sich mithilfe des APSC tiefere Einblicke in physiologische Zusammenhänge (mögliche off-targets oder weitere on-targets) erzielen, welche unter anderem für das Design der nachfolgenden Experimente wertvoll sind. Andererseits bietet der MSS-Ansatz die Möglichkeit die Ähnlichkeit auf verschiedenen Ebenen zu betrachten und in Kombination mit weiteren Aktivitätsdaten die für das Zielprotein am besten geeignete Ebene zur Optimierung von Liganden einer spezifischen Zielstruktur zu ermitteln.



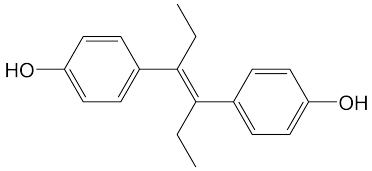
(a) ChEMBL246437
$K_i = 106$ nM



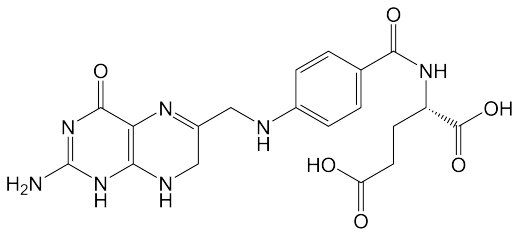
(b) ChEMBL399601
$K_i = 12500$ nM



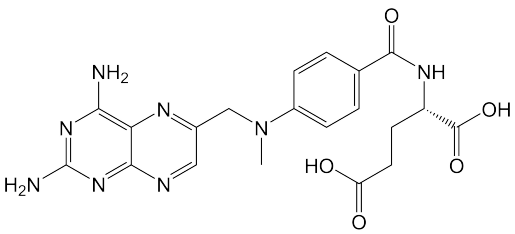
(c) Estradiol
$K_i = 0{,}113$ nM



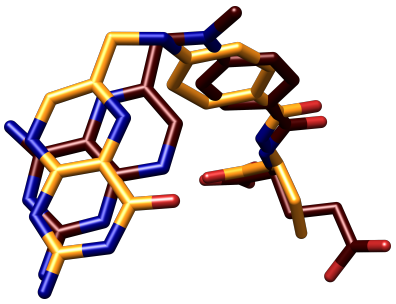
(d) Diethylstilbestrol
$K_i = 0{,}26$ nM



(e) Dihydrofolsäure



(f) Methotrexat



(g) Proteinstruktur-basiertes Alignment von Dihydrofolsäure und Methotrexat

Abbildung I.0.1.: Beispiele für molekulare Ähnlichkeit mit biologisch verschiedener Aktivität gegenüber dem Dopamin D3-Rezeptor (a) und (b) (Dimova *et al.* 2013) und Bioähnlichkeit bezüglich des Estrogenrezeptors α trotz distinkter Struktur (c) und (d). Die Verbindungen Dihydrofolsäure (e) und Methotrexat (f) interagieren mit der Dihydrofolatreduktase, wobei das anellierte Ringsystem nicht aligniert ist (g). Die Aktivitätsdaten entstammen der ChEMBL. Für das Proteinstruktur-basierte Alignment (g) wurden die PDB-IDs: 1RF7 (orange, Dihydrofolsäure) und 1RA3 (dunkelrot, Methotrexat) verwendet.

I.2.1. Der Maximum Similar Subgraph Ansatz (MSS-Ansatz)

Beim MSS-Ansatz erfolgt eine Kombination einer Graph-basierten Ähnlichkeitsanalyse mit Vorgaben zur Abbildung der Bioähnlichkeit. Die Ähnlichkeitsanalyse erfolgt entweder durch eine Ähnlichkeitssuche (BBPMCS-Algorithmus, für Details siehe Abschnitt II.1.3 und Droschinsky, Kriege und Mutzel (2017)) oder mittels eines Clustering-Algorithmus (StruClus, für Details siehe Abschnitt II.1.3 und Schäfer und Mutzel (2017)). Die abgebildeten Aspekte der Bioähnlichkeit betreffen die Definition von bioähnlichen Atomtypen, pharmakophoren Eigenschaften, bioisosteren Paaren und Shape (Form).

I.2.2. Das Advanced Privileged Scaffold Concept (APSC)

Eine Übertragung einer Ligand-basierten Ähnlichkeit auf eine Ähnlichkeit der entsprechenden Targetproteine wurde im APSC umgesetzt. Dazu wird für die Liganden basierend auf einem gemeinsamen Grundgerüst (Scaffold) eine Ähnlichkeit definiert (Scaffold-Ebene). Sind die Targetproteine der Liganden weder Sequenz-ähnlich noch funktionell verwandt, so liegt ein privilegiertes Scaffold vor. Eine Ähnlichkeit der Proteine basierend auf privilegierten Strukturen zur Erkennung und Bindung des privilegierten Scaffolds kann folglich angenommen werden. Kann diese Annahme z. B. durch die Übereinanderlagerung der Kristallstrukturen oder durch Bindetaschenähnlichkeiten bestätigt werden, handelt es sich um erweiterte privilegierte Scaffolds. Dies kann zur Entdeckung neuer Liganden, selektiver Liganden oder Multitargetliganden (Polypharmakologie) ausgenutzt werden. Eine beispielhafte Veranschaulichung des APSC mit Relevanz für die vorliegende Arbeit befindet sich in der Abbildung IV.2.2.

I.3. Aufgabenstellung

Im Rahmen dieser Arbeit sollte die Ähnlichkeit von niedermolekularen Verbindungen unter Berücksichtigung der biologischen Aktivität betrachtet werden. Dazu sollte zunächst eine Datenbank solcher Verbindungen generiert werden, die einen weiten chemischen Raum abdeckt und deren Verbindungen größtenteils zugleich synthetisch zugänglich sind. Diese sollte als Ideengenerator in zukünftigen Wirkstoffdesignprojekten verwendet werden können, indem die in den nachfolgenden Analysen zur Ähnlichkeit entwickelten Methoden auf sie angewendet werden. Anschließend sollten mehrere Ebenen der Ähnlichkeit von niedermolekularen Verbindungen untersucht werden: (i) Atomtypengruppen und pharmakophore Eigenschaften, (ii) Scaffolds (Grundgerüste) sowie (iii) Bioisostere und Shape (Form). Dabei sollten zwei unterschiedliche Methoden entwickelt werden. Die erste Methode sollte die gewonnenen Erkenntnisse aus den Ähnlichkeitsanalysen in Graph-basierte Such- oder Clustering-Algorithmen zur Erfassung der Bioähnlichkeit von niedermolekularen Verbindungen integrieren. Die zweite Methode sollte eine Erfassung der biologischen Zielstruktur (Protein) in der Betrachtung der Bioähnlichkeit von niedermolekularen Verbindungen erlauben. Zur Validierung sollten die Vorhersagen dieser Methode biochemisch verifiziert werden. Die entwickelten Methoden sollen eine Einbeziehung der biologischen Aktivität in die Ähnlichkeitsbetrachtung von niedermolekularen Verbindungen erlauben und somit zusammen mit der generierten Datenbank die Wirkstoffentwicklung fördern.

Teil II.

Grundlagen

II.1. Repräsentation von Molekülen und Ähnlichkeitsfunktionen

Ein erster essentieller Schritt in der computer-gestützten Wirkstoffforschung (u. a. zur Durchführung von Ähnlichkeitsanalysen) ist, Moleküle in eine Repräsentation zu überführen, welche zur Verarbeitung durch Algorithmen geeignet ist. Dieser Schritt, auch wenn häufig wenig beachtet, ist eine entscheidende Grundlage für alle weiterführenden Analysen und Methoden, denn nur Eigenschaften, die vom Molekül in die Repräsentation überführt werden, können in den nachfolgenden Untersuchungen berücksichtigt werden. Damit bestimmt die gewählte Repräsentation maßgeblich über die Art der Ähnlichkeit, welche betrachtet wird. Im Folgenden sollen einerseits eine häufig verwendete Repräsentation (Fingerprints) und andererseits eine komplexere Repräsentation (Graphen) vorgestellt werden. Zudem kann die Ähnlichkeit anhand eines speziellen Teils des Moleküls (z. B. Scaffold oder Fragment bzw. Substruktur wie z. B. eine funktionelle Gruppe), mehreren Teilen oder anhand der vollständigen Molekülstruktur ermittelt werden. Ein weiterer essentieller Punkt mit Auswirkung auf die Definition der Ähnlichkeit und die Art der Ähnlichkeit welche untersucht wird, ist die Wahl einer geeigneten Ähnlichkeitsfunktion. Die Ähnlichkeitsfunktion bewertet die Ähnlichkeit zweier Molekül-Repräsentationen und besitzt somit ebenfalls einen entscheidenden Einfluss auf die Definition der Ähnlichkeit. Sie versucht, die subjektive Wahrnehmung von Ähnlichkeit zu rationalisieren und zu quantifizieren. Dazu wird die Ähnlichkeit mithilfe eines objektiven sowie einfach zu vergleichenden numerischen Wertes ausgedrückt.

II.1.1. Fingerprints

Die Molekülstruktur und Eigenschaften eines Moleküls können als Strings bzw. Vektoren repräsentiert werden. Diese Strings werden Fingerprints genannt. Ursprünglich wurden sie häufig als Vorfilterung für die Substruktursuche verwendet. Fingerprints haben sich jedoch auch für die Ähnlichkeitssuche als sehr nützlich erwiesen (Leach und Gillet 2007). Ein Vergleich von Fingerprints ist mit wenig Rechenaufwand möglich, weshalb sie ebenfalls für Substruktur- und Ähnlichkeitssuchen in großen Datenbanken verwendet werden können. Soll hingegen die Ähnlichkeit eines Datensatzes zu einem anderen mittels Fingerprints ermittelt werden, wird ein paarweiser Vergleich aller Strukturen benötigt. Dies ist wenig effizient und nur bedingt für große Datensätze realisierbar. Zudem stellen Fingerprints eine Vereinfachung der Molekülstruktur dar, die mit einem Informationsverlust einhergeht. Im Folgenden sollen zwei häufig verwendete Fingerprint-Arten zur Repräsentation von niedermolekularen Verbindungen (Fragment-basierte und Hash-basierte Fingerprints) und die Klasse der Protein-Ligand-Interaktions-Fingerprints vorgestellt werden. Für eine umfassende Übersicht der Fingerprints und ihrer Performanz sei auf Riniker und Landrum (2013) und O'Boyle und Sayle (2016) verwiesen.

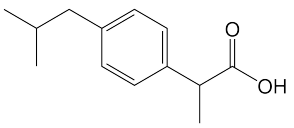
Außerdem können Fingerprints in binäre Fingerprints, welche lediglich An- (1) oder Abwesenheit (0) eines Merkmals des Moleküls anzeigen (schematische Darstellung am Beispiel Ibuprofen siehe Abbildung II.0.1c) und nicht-binäre Fingerprints (z. B. *count* Fingerprints, bei denen die Anzahl, wie oft ein Merkmal, z. B. eine bestimmte Substruktur vorkommt, gezählt wird) unterteilt werden. Zu beachten ist, dass mit dem Anstieg der Größe oder Komplexität eines Moleküls der Anteil an gesetzten Bits (z. B. Anteil an Einsen) in seiner Fingerprintrepräsentation zunimmt (Maggiora *et al.* 2014). Dies führt zu einer statistischen Erhöhung der Wahrscheinlichkeit, einen höheren Ähnlichkeitswert für größere oder komplexere Moleküle mittels einer nicht entsprechend normalisierten Ähnlichkeitsfunktion zu erzielen.

Fragment-basierte Fingerprints

Bei Fragment-basierten Fingerprints wird das Molekül nach Fragmenten (bzw. Substrukturen) durchsucht, welche in einer vordefinierten Liste aufgeführt sind. Jedem Fragment ist eine spezifische Position im Fingerprint zugeordnet. Kann das Fragment in dem Molekül gefunden werden, wird die dem Fragment entsprechende Position im Bit-String auf eins gesetzt. Eine schematische Darstellung am Beispiel von Ibuprofen befindet sich in der Abbildung II.0.1c. So generierte Fingerprints sind einfach nachvollziehbar. Von Nachteil kann sein, dass vordefinierte Fragmente verwendet werden. Besonderheiten eines Moleküls können so nicht immer adäquat abgebildet werden und verschiedene Moleküle mit gleichen an- und abwesenden Fragmenten aus der vordefinierten Liste können nicht unterschieden werden. Ein Beispiel aus der Klasse der Fragment-basierten Fingerprints ist der MACCS-Key, welcher 166 Fragmente umfasst und gut für die Vorfilterung im Rahmen von Substruktursuchen geeignet ist. In Abbildung II.0.1d wird als Beispiel der MACCS-Key für Ibuprofen gezeigt.

Hash-basierte Fingerprints

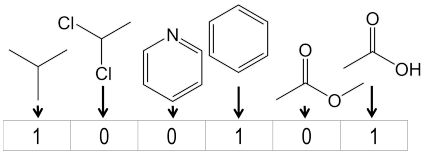
Hash-basierte Fingerprints generieren die Fragmente während der Erstellung des Fingerprints (*on the fly*). Dadurch sind sie nicht auf vordefinierte Fragmente angewiesen und können somit spezielle Eigenschaften des Moleküls oder des Datensatzes besser abbilden. Den generierten Fragmenten werden mittels einer Hash-Funktion ganzzahlige Werte zugewiesen. Diese definieren die Position im Fingerprint, welche bei Anwesenheit moduliert wird. Oft werden mehrere Hash-Funktionen verwendet und somit für ein Fragment mehrere Positionen moduliert. Zu beachten ist, dass die Hash-Funktion in der Regel keine eineindeutige Zuordnung ist, dies bedeutet, dass für verschiedene Fragmente die gleiche Position im Fingerprint moduliert werden kann. Man spricht in diesem Fall von einer *Kollision*. Ein Hash-basierter Fingerprint, welcher mit dem Ziel entwickelt wurde, besonders Fragmente mit Bedeutung für die Aktivität zu berücksichtigen, ist der *extended-connectivity fingerprint 4* (ECFP4) (Rogers und Hahn 2010). Dieser ist beispielhaft für Ibuprofen in Abbildung II.0.1b gezeigt.



(a) Ibuprofen

734603939 -1910270391 1559650422 642810091
-182236392 -1100000244 -1074141656 -1884411803
864518973 1035165602 770157610 -2024255407
1996767644 -176846085 737860717 -1905455774
2099970318 2025485523 -743156170 462865601
1799026729 1088861418 -395008465 -938504044
-484332638 395448388

(b) ECFP4



(c) schematische Darstellung

74 115 123 139 141 149 154 155 157 159
160 162 163 164 165

(d) MACCS-Key

Abbildung II.0.1.: Ibuprofen (a), Hash-basierte Fingerprintrepräsentation von Ibuprofen am Beispiel vom ECFP4 (b), schematische Darstellung eines binären Fragment-basierten Fingerprints für Ibuprofen (c) und Fragment-basierte Fingerprintrepräsentation von Ibuprofen am Beispiel des MACCS-Keys (d). Der Pfeil gibt an, welche Position im darunter abgebildeten Fingerprint dem darüber abgebildeten Fragment zugeordnet ist.

Der ECFP4 wird häufig für Ähnlichkeitssuchen und virtuelle Screeningansätze verwendet. In virtuellen Screenings zeichnet sich der ECFP4 besonders durch eine große Fläche unter der *Receiver-Operating-Characteristic* (ROC) Kurve (*area under the curve* (Fläche unter der Kurve) (AUC)) aus (Riniker und Landrum 2013). Die AUC der ROC Kurve ist ein Maß für die Performanz z. B. einer virtuellen Screeningmethode. Der Wertebereich liegt zwischen null und eins, wobei Werte über 0,5 eine im Vergleich zum Zufall bessere Methode implizieren. In dieser Arbeit wird der ECFP4 ebenfalls für Ähnlichkeitssuchen und als State of the Art molekularer Fingerprint zur Bewertung der neu entwickelten Methoden verwendet (Duvenaud *et al.* 2015). Eine Beschreibung des Algorithmus befindet sich in Rogers und Hahn (2010) und ein Vergleich mit dem MACCS-Key wurde von Maggiora *et al.* (2014) durchgeführt.

Protein-Ligand-Interaktions-Fingerprints

Im Bezug auf die Einbindung von Eigenschaften der Moleküle, die für die Aktivität wichtig sind, sind neben den Hash-basierten Fingerprints auch die Protein-Ligand-Interaktions-Fingerprints von Bedeutung.

Im Gegensatz zu den Hash-basierten Fingerprints wird der Fokus bei der Repräsentation nicht auf die gesamte Molekülstruktur, sondern nur auf die Bereiche (Substrukturen), welche mit der Zielstruktur (dem Protein) interagieren, gelegt. Dadurch wird eine auf die Interaktion mit einem Protein spezialisierte Repräsentation ermöglicht, die insbesondere in virtuellen Screeningansätzen eine gute Performanz aufweist. Eine Übersicht der Protein-Ligand-Interaktions-Fingerprints wird in Rognan und Desaphy (2013) gegeben. Ein kürzlich entwicklter Protein-Ligand-Interaktions-Fingerprint ist PADIF (*protein per atom score contributions derived interaction fingerprint*) (Jasper *et al.* 2018). PADIF wertet die beim Docking der niedermolekularen Verbindung in die Proteinbindetasche mittels GOLD (Jones *et al.* 1997) generierten Atomscores neu aus und gewichtet sie. Im Gegensatz zu den beiden anderen vorgestellten Fingerprints wird für die Generierung des PADIFs eine Proteinstruktur und die Vorprozessierung mittels Docking benötigt. Dadurch kann der PADIF besser Protein-spezifische Ähnlichkeiten abbilden.

Als nächstes soll die Repräsentation von Molekülen als Graphen vorgestellt werden. Dazu ist jedoch zuvor ein kleiner Exkurs in die Graphtheorie sinnvoll.

II.1.2. Exkurs: Graphtheorie

Definition 1 (Graph). *Ein* Graph $G = (V, E)$ *besteht aus einer endlichen Menge an* Knoten *(V) und* Kanten *(E). Dabei wird $V(G) := V$ und $E(G) := E$ definiert. Die Kanten verbinden dabei die Knoten.*

Ein Knoten $u \in V$ ist zu einem Knoten $v \in V$ *adjazent*, falls die Kante $\{u, v\} \in E$ existiert. Der *Grad* eines Knotens v ist die Anzahl der zu v adjazenten Knoten.

Definition 2 (gelabelter Graph). *Ein Graph ist* gelabelt*, wenn seine Kanten und/oder Knoten auf ein endliches Alphabet (endliche Menge) abgebildet werden.*

In dieser Arbeit werden, sofern nicht anders angeben, ungerichtete gelabelte Graphen mit endlicher Kanten- und Knotenmenge verwendet. Ein Beispiel anhand von Ibuprofen mit Knotenlabeln entsprechend des Elements befindet sich in Abbildung II.0.2b.

Definition 3 (vollständiger Graph). *Ein Graph $G = (V, E)$ heißt vollständig, falls $E = \{\{u, v\} \mid u, v \in V, u \neq v\}$, also alle Knoten paarweise miteinander durch eine Kante verbunden sind.*

Definition 4 (Pfad). *Ein* Pfad $P = (v_1, v_2, \ldots, v_k)$ *in einem Graphen $G = (V, E)$ ist eine Abfolge von Knoten aus V, welche durch Kanten ($v_i, v_i + 1 \in E$ für alle $1 \leq i < k$) miteinander verbunden sind. P beginnt und endet mit einem Knoten. v_1 wird dabei* Startknoten *und v_k wird* Endknoten *des Pfades genannt.*

Definition 5 (Kreis). *Ist nur der Start- und Endknoten eines Pfades identisch, so wird von einem* Kreis *gesprochen.*

Definition 6 (zusammenhängender Graph). *Existiert zwischen jedem Knotenpaar u und v in G $(u, v \in V)$ mindestens ein Pfad P, so ist G* zusammenhängend *(connected).*

Definition 7 (Subgraph). *Ein* Subgraph *(Teilgraph) von G ist ein Graph $G' = (V', E')$ mit $V' \subseteq V$ und $E' \subseteq E$. Der so definierte Subgraph kann auch als $G' \subseteq G$ geschrieben werden.*

Definition 8 (Zusammenhangskomponente). *Die maximalen zusammenhängenden Subgraphen von G nennt man Zusammenhangskomponenten. Ein zusammenhängender Graph hat dementsprechend eine Zusammenhangskomponente.*

Definition 9 (Brücke). *Führt die Entfernung der Kante $e \in E$ dazu, dass sich die Anzahl an Zusammenhangskomponenten in G erhöht, so wird e eine Brücke genannt.*

Definition 10 (zweizusammenhängender Graph). *Ein Graph $G = (V, E)$ ist* zweizusammenhängend *(biconnected), wenn beim Entfernen eines beliebigen Knotens $v \in V$ der daraus entstehende Graph $G \setminus \{v\}$ weiterhin zusammenhängend ist.*

Definition 11 (Block). *Ein* Block *ist ein maximaler zweizusammenhängender Subgraph (Beispiel anhand von Ibuprofen siehe Abbildung II.0.2c).*

Definition 12 (Clique). *Eine* Clique *ist ein (Sub-)Graph, zwischen dessen Knoten jeweils eine verbindende Kante existiert, er ist folglich* vollständig.

Definition 13 (planarer Graph). *Ein Graph ist* planar*, wenn eine Zeichnung des Graphen in der Ebene möglich ist, ohne dass sich Kanten kreuzen.*

In einer kreuzungsfreien Zeichnung eines planaren zusammenhängenden Graphen G in der Ebene wird die durch die Kanten und Knoten eingeschlossene zusammenhängende Fläche von der uneingeschränkten *äußeren Fläche* unterschieden.

Definition 14 (außenplanarer Graph). *Existiert eine kreuzungsfreie Zeichnung für einen planaren Graphen G, sodass jeder Knoten an der äußeren Fläche liegt, so ist der Graph* außenplanar*. In diesem Fall kann folglich jeder Knoten von außen erreicht werden, ohne eine Kante zu kreuzen.*

Definition 15 (reduzierter Graph). *Werden in einem Graphen aneinandergrenzende (adjazente) Knoten zusammengefasst (Kantenkontraktion), so spricht man beim resultierenden Graphen von einem* reduzierten Graph *(Beispiele siehe Abbildungen II.0.2d, II.0.2e und II.0.2f).*

Definition 16 (Baum). *Ein* Baum *T ist ein zusammenhängender Graph, der keinen Kreis enthält.*

Isomorphismen

Definition 17 (Isomorphismus). *Ein* Isomorphismus *zwischen zwei Graphen G und H ist eine Bijektion $\psi : V(G) \to V(H)$, sodass $\{u, v\} \in E(G) \iff \{\psi(u), \psi(v)\} \in E(H)$ für alle $u, v \in V(G)$. Falls ein Isomorphismus zwischen den Graphen G und H existiert, werden sie als* isomorph *($G \simeq H$) bezeichnet.*

Definition 18 (Subgraphisomorphismus). *Ein* Subgraph-Isomorphismus *von einem Graphen G zu einem Graphen H ist ein Isomorphismus zwischen G und einem Subgraphen $H' \subseteq H$.*

Definition 19 (gemeinsamer Subgraphisomorphismus). *Ein* gemeinsamer Subgraph-Isomorphismus *zwischen G und H ist ein Isomorphismus zwischen zusammenhängenden Subgraphen $G' \subseteq G$ und $H' \subseteq H$.*

Definition 20 (block and bridge preserving Subgraph). *Ein Subgraph G' ist block and bridge preserving* (Block und Brücken beibehaltend) (BBP) *falls (i) jede Brücke in G' eine Brücke in G ist und (ii) jede zwei Kanten in verschiedenen Blöcken in G' in verschiedenen Blöcken in G sind.*

Definition 21 (Schnittknoten). *Zerfällt ein Graph durch die Entfernung eines Knotens $v \in V$ in mehrere Zusammenhangskomponenten (ist er nicht mehr zusammenhängend), so wird v als* Schnittknoten *bezeichnet.*

Definition 22 (Gewichtsfunktion). *Eine Gewichtsfunktion $w : V \times V \cup E \times E \to \mathbb{R}$ kann für die Abbildung von Knoten auf Knoten (Abbildung verschiedener z. B. Atomtypen aufeinander) und Kanten auf Kanten (Abbildung verschiedener Bindungstypen aufeinander) verwendet werden.*

Definition 23 (Gewicht). *Das Gewicht $W(\psi)$ eines Isomorphismus ψ zwischen G und H unter der Gewichtsfunktion w ist die Summe der Gewichte $w(v, \psi(v))$ und $w(\{u, v\}, \{\psi(v), \psi(v)\})$ für alle Knoten v und Kanten $\{u, v\}$, die durch ψ abgebildet werden konnten.*

Definition 24 (größter gemeinsamer Subgraph). *Ein gemeinsamer Subgraph-Isomorphismus ist* am größten*, man spricht auch vom maximum common subgraph (größter gemeinsamer Teilgraph) (MCS), wenn sein Gewicht $W(\psi)$ am größten ist.*

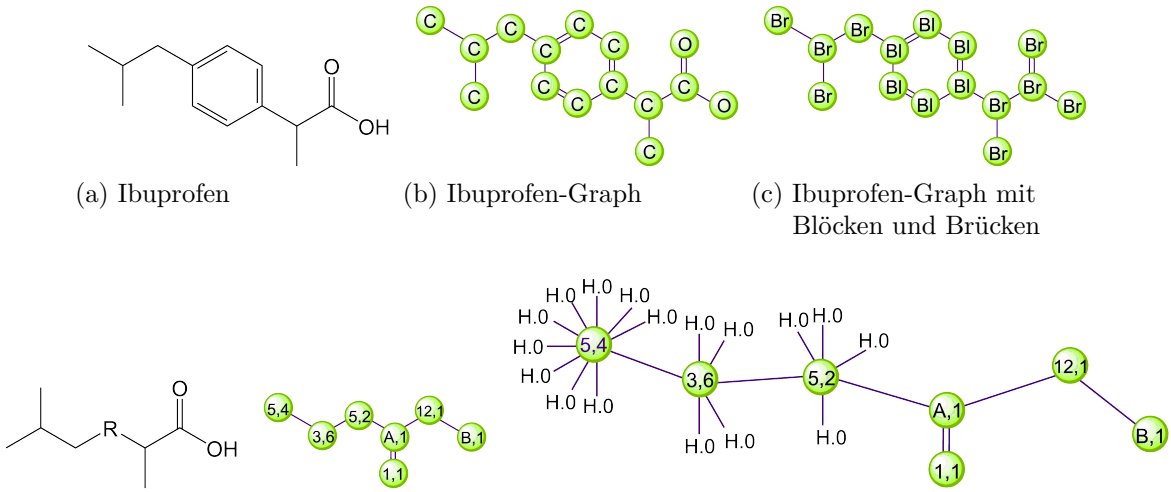
Nach diesem Exkurs wird nun die Repräsentation von Molekülen als Graphen behandelt.

II.1.3. Repräsentation von Molekülen als Graphen - Molekülgraphen

Moleküle können als Graphen repräsentiert werden (Beispiel anhand von Ibuprofen siehe Abbildung II.0.2b). Dadurch ist eine genauere Repräsentation des Moleküls möglich und es konnte gezeigt werden, dass Fingerprint-basierte und Graph-basierte Repräsentationen unterschiedliche Ähnlichkeiten detektieren können (Raymond und Willett 2002). Bei der Graph-basierten Repräsentation können die Atome den Knoten und die Bindungen den Kanten entsprechen. Dementsprechend werden die Kanten mit Bindungsart-Labeln und die Knoten mit Atomtyp-Labeln versehen. Diese Label sollen während eines Isomorphismus erhalten bleiben. Die Abbildung verschiedener Label aufeinander kann mithilfe einer Gewichtungsfunktion w reglementiert werden. Ein gemeinsamer Subgraph mehrerer Molekülgraphen entspricht einer gemeinsamen Substruktur dieser.

Der Maximum Common Subgraph (MCS) und der Maximum Similar Subgraph (MSS) in Bezug auf Moleküle

Der *MCS* wird häufig zur Ermittlung der Ähnlichkeit von Molekülen verwendet (Kriege, Humbeck und Koch 2019) und entspricht der größten gemeinsamen Substruktur dieser. Jedoch sind bei dieser Methode nur Moleküle ähnlich, welche eine identische Substruktur besitzen. Bezogen auf die Ähnlichkeit hinsichtlich gleicher oder ähnliche biologischer Wirksamkeit/Aktivität ist diese Herangehensweise oft zu restriktiv. So ist es häufig ausreichend die pharmakophore Eigenschaft, die Gestalt bzw. Form, das Interaktionsmuster oder die Anordnung von physikochemischen Eigenschaften beizubehalten, unabhäning von der exakt gleichen Anordnung derselben Elemente mit denselben Bindungen.
Aus diesem Grund wird in dieser Arbeit der *MSS* eingeführt. Das Ziel des MSS ist, diejenigen Differenzen zwischen Subgraphen zu tolerieren, welche mit hoher Wahrscheinlichkeit nicht zur Veränderung der biologischen Aktivität führen. Ein weiteres Problem besteht darin, dass die Berechnung eines MCS für allgemeine Graphen NP-schwer (vermutlich nicht in Polynomialzeit lösbar) ist. Damit ist die Laufzeit im Worst Case exponentiell zur Graphgröße. Allerdings besitzen Molekülgraphen in der Regel eine geringe Graphgröße und weitere besondere Eigenschaften, die man zur Laufzeitreduktion nutzen kann. Aus diesem Grund werden auf Moleküle angepasste Algorithmen entwickelt, die z. B. auf spezielle Graphklassen zugeschnitten sind oder von Heuristiken profitieren und damit quadratische oder lineare Laufzeiten im Bezug auf die Anzahl an zu untersuchenden Molekülgraphen besitzen können. Zwei solche Algorithmen für die Suche und das Clustering von Molekülen sind der *block and bridge preserving maximum common subgraph* (BBP größter gemeinsamer Teilgraph) (BBPMCS) (siehe Abschnitt II.1.3) und StruClus (siehe Abschnitt II.1.3).



(a) Ibuprofen (b) Ibuprofen-Graph (c) Ibuprofen-Graph mit Blöcken und Brücken

(d) Ibuprofen mit reduziertem Benzenring (e) reduzierter Ibuprofen-Graph ohne Wasserstoffatome (f) reduzierter Ibuprofen-Graph mit Wasserstoffatomen

Abbildung II.0.2.: Ibuprofen (a), Graphrepräsentation von Ibuprofen mit Knotenlabel entsprechend des Elements (b), Graphrepräsentation von Ibuprofen mit Knotenlabel entsprechend der Zuordnung zu einer Brücke (bezieht sich auf die zu Br adjazenten Kanten) oder einem Block (c), Ibuprofen-Struktur mit zu R reduziertem Benzenring (d), reduzierter Ibuprofen-Graph mit Knotenlabeln entsprechend der Atomtypengruppe (Zeichen vor dem Komma) und der Gruppengröße (Zeichen nach dem Komma) ohne Berücksichtigung von Wasserstoffatomen (e), reduzierter Ibuprofen-Graph mit Knotenlabeln entsprechend der Atomtypengruppe (Zeichen vor dem Komma) und der Gruppengröße (Zeichen nach dem Komma) unter Berücksichtigung von Wasserstoffatomen (f) (Kanten in lila, Knoten in grün, Br: Brücke, Bl: Block, A: C.co2h, B: H.ac).

Der StruClus-Algorithmus

StruClus (Schäfer und Mutzel 2017) ist ein struktureller Graphen-basierter Clustering-Algorithmus, welcher einen Moleküldatensatz iterativ anhand von häufigen Substrukturen, welche *on the fly* berechnet werden, in Cluster einteilt. Diese häufigen Substrukturen dienen als Repräsentanten des Clusters. Eine besondere Eigenschaft dieses Clustering-Algorithmus ist, dass nicht nur ein Graph als Repräsentant dient, sondern mehrere solcher Subgraphen einen Cluster repräsentieren. Dies ermöglicht leichte Variationen innerhalb eines Clusters abzudecken, um z. B. die Bioähnlichkeit besser abzubilden. So können z. B. verschiedene Moleküle, welche mittels verschiedener, jedoch physikochemisch ähnlichen, Substrukturen mit demselben Proteintarget interagieren in einem gemeinsamen Cluster enthalten sein, welcher von den verschiedenen, jedoch physikochemisch ähnlichen Gruppen repräsentiert wird. Damit dienen die Repräsentanten als intuitive Beschreibung des jeweiligen Clusters. Eine weitere wichtige Eigenschaft von StruClus ist, dass der Algorithmus mit Noise (Rauschen) umgehen kann. Dies ist insbesondere bei der Chemieinformatik wichtig, da die Zuverlässigkeit bestimmter Angaben wie Bioaktivitäten nicht immer gewährleistet ist. Zudem verhält sich die Laufzeit linear zur Menge der Eingabegraphen bzw. Eingabemoleküle (Schäfer und Mutzel 2017). Damit ist StruClus auch für Moleküldatenbanken mit hunderten Millionen Molekülen, wie CH*I*PMUNK (Humbeck *et al.* 2018), anwendbar. Dies macht StruClus besonders attraktiv für die stets anwachsenden Moleküldatenbanken.

Der Block and Bridge Preserving Maximum Common Subgraph Algorithmus - BBPMCS

Der BBPMCS durchsucht einen Datensatz nach einer Suchstruktur oder einer Liste von Suchstrukturen und gibt die jeweilige Ähnlichkeit aus. Zudem wird ermittelt, welche Atome (Knoten) der Moleküle (Graphen) aufeinander abgebildet (gemappet) werden können. Durch die Zerteilung der zu vergleichenden Graphen G und H in Brücken und Blöcke sowie verbindende Schnittknoten, hat der Algorithmus von Droschinsky, Kriege und Mutzel (2017) eine quadratische Laufzeit $\mathcal{O}(|G||H|)$, sofern die zu vergleichenden Blöcke außenplanar sind und der Grad der Knoten klein ist. Dies ist für Moleküle meistens der Fall. Ein Kohlenstoffatom hat einen Grad von in der Regel maximal vier. Die meisten Molekülgraphen sind zudem außenplanar und erfüllen damit die letzte Bedingung (Horváth, Ramon und Wrobel 2010). Die Zuordnung der Brücken in G und H erfolgt effizient über den größten gemeinsamen Teilbaum Isomorphiealgorithmus von Droschinsky, Kriege und Mutzel 2016. Es können acht verschiedene Ähnlichkeitsfunktionen verwendet werden. Falls nicht anders im Text erwähnt, wird die Funktion nach Wallis *et al.* (2001) (ID 1, siehe Tabelle III.2.4) verwendet. Eine Übersicht der Ähnlichkeitsfunktionen inklusive Analyse befindet sich im Abschnitt III.2.2.3.

Im Bezug auf die Analyse der Ähnlichkeiten von niedermolekularen Verbindungen hat der Algorithmus von Droschinsky, Kriege und Mutzel (2017) den Vorteil, dass es häufig gewünscht ist, Ringe (entsprechen Blöcken) auf diese abzubilden und auch Linker (entsprechen Brücken) nur auf andere Linker bzw. Brücken abzubilden. Somit erhält Block and Bridge Preserving die lokale Konnektivität der Eingabegraphen (Droschinsky, Kriege und Mutzel 2017). Da eine Gewichtsfunktion übergeben werden kann, ist es zudem möglich, Ähnlichkeiten zu definieren bzw. verschieden stark zu gewichten. In der vorliegenden Arbeit wurden drei verschiedene Gewichtsfunktionen (Ähnlichkeitsmatrizen) definiert (siehe Kapitel IV.2.11.2).

II.1.4. Ähnlichkeitsfunktionen

Zum rationalen Vergleich zweier Molekül-Repräsentationen wird eine Ähnlichkeitsfunktion benötigt. Eine Überischt zu gebräuchlichen Ähnlichkeitsfunktionen und deren Performanz sowohl für eine Fingerprint-basierte als auch für eine Graph-basierte Repräsentation von Molekülen befindet sich in Raymond und Willett (2002). In der Chemieinformatik ist der Jaccard-Tanimoto-Koeffizient (kurz Tanimoto-Koeffizient) (TC) zum Vergleich von Fingerprints weitverbreitet. Der TC wird nach Gleichung II.0.1 berechnet, wobei A die Menge an vorhandenen Eigenschaften im Fingerprint a ist und B die Menge an vorhandenen Eigenschaften im Fingerprint b ist.

$$TC(a,b) = \frac{|A \cap B|}{|A \cup B|} \tag{II.0.1}$$

In dieser Arbeit wird neben dem TC auch der Dice-Koeffizient (DC) zur Bewertung der Ähnlichkeit von Fingerprint-basierten Molekül-Repräsentationen verwendet. Die Berechnung des DC ist gegeben in Gleichung II.0.2, wobei A wieder die Menge an vorhandenen Eigenschaften im Fingerprint a ist und B die Menge an vorhandenen Eigenschaften im Fingerprint b ist. Es konnte gezeigt werden, dass sowohl der TC als auch der DC eine gute Performanz in Ähnlichkeitssuchen aufweisen (Bajusz, Rácz und Héberger 2015). Beide vorgestellten Koeffizienten besitzen einen Wertebereich von null bis eins, wobei ein Wert von eins die maximale Ähnlichkeit angibt (dies muss jedoch nicht bedeuten, dass die Moleküle identisch sind).

$$DC(a,b) = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{II.0.2}$$

Die in dieser Arbeit verwendeten Graph-basierten Ähnlichkeitsfunktionen befinden sich in der Tabelle III.2.4.

Zusammenfassend gibt es weder die eine richtige Repräsentation eines Moleküls noch eine einzig richtige Ähnlichkeitsfunktion. Beide Faktoren sind abhängig von der Art der Ähnlichkeit, die betrachtet und analysiert werden soll, sowie dem damit verfolgten Ziel.

Eine Kritik am TC besteht darin, dass die TC-Werteverteilung auch bei zufälligen Datensätzen Häufungen (Peaks) bei definierten Werten hat (z. B. 1/3) (Godden, Xue und Bajorath 2000). In dieser Arbeit konnte ebenfalls beobachtet werden, dass die TC-Werteverteilung bei niedrigen TC-Werten in der vorliegenden Analyse im Vergleich zu der Werteverteilung der neu entwickelten Methode (MSS) zu einem höheren Wert verschoben ist (siehe Abbildung A.2). Dies bedeutet, dass eine höhere Grundähnlichkeit angenommen wird bzw. das Rauschen beim TC stärker ausgeprägt ist. Dieser Effekt ist besonders für Diversitätsanalysen von Bedeutung, da dabei kleine TC-Werte betrachtet werden und in diesem Bereich vermehrt zufällige Häufungen auftreten.

II.2. BRD4 und PPARγ

II.2.1. BRD4

BRD4 gehört zu den Bromodomain-containing Proteinen (BRDs), welche in 8 Familien eingeteilt werden und insgesamt 61 Bromodomänen sowie 46 Proteine umfassen (Filippakopoulos und Knapp 2014). BRD4 gehört zur Familie der *bromodomain and extra terminal domain* (BET) Proteine der Familie II (Filippakopoulos und Knapp 2014; Ferri, Petosa und McKenna 2016). Diese umfasst die Proteine BRD2, BRD3, BRD4 und BRDT, welche alle zwei Bromodomänen (BD1 und BD2) und eine extra-terminale Domäne (ET) enthalten (Shi und Vakoc 2014). BRD4 enthält zusätzlich noch eine C-terminale Domäne (CTD) (siehe Abbildung II.0.3).

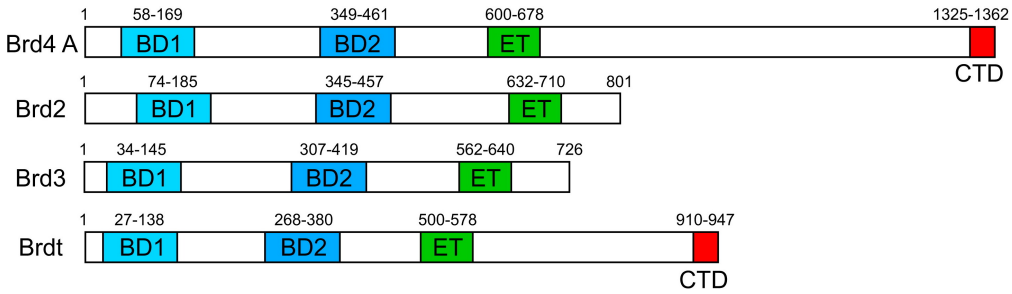


Abbildung II.0.3.: Aufbau der Proteine der BET-Familie. BD1: Bomodomäne 1, BD2: Bromodomäne 2, ET: extra-terminale Domäne, CTD: C-terminale Domäne. Verändert nach Shi und Vakoc (2014).

Die Bromodomänen sind aus vier α-Helices und verbindenden Loopregionen aufgebaut, welche eine hydrophobe Tasche bilden (Shi und Vakoc 2014). In dieser Tasche können ε-N-acetylierte Lysine (Kac) binden. Alle Proteine der BET-Familie können diacteylierte Peptide gleichzeitig innerhalb einer Bromodomäne erkennen (Ferri, Petosa und McKenna 2016). Acetylierte Lysine spielen eine wichtige Rolle bei Protein-Protein-Interaktionen und im epigenetischen Code (Filippakopoulos und Knapp 2014).

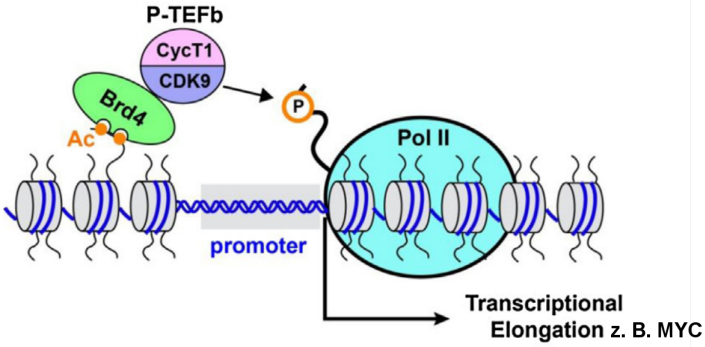
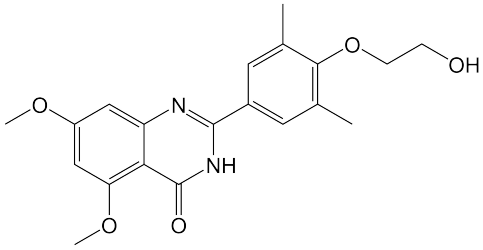
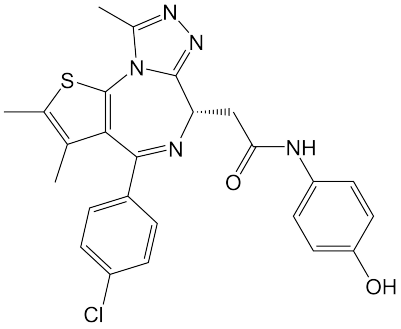


Abbildung II.0.4.: Physiologische Funktion von BRD4 als Reader von acetylierten Proteinen und Transkriptionsregulator. P-TEFb: positiver Transkriptionselongationsfaktor bestehend aus Cyclin T1 (CycT1) und Cyclin-abhängiger Kinase 9 (CDK9), Pol II: RNA-Polymerase II, Ac: acetylierte Aminosäuren, P: Phosphorylierung. Verändert nach Ferri, Petosa und McKenna (2016).

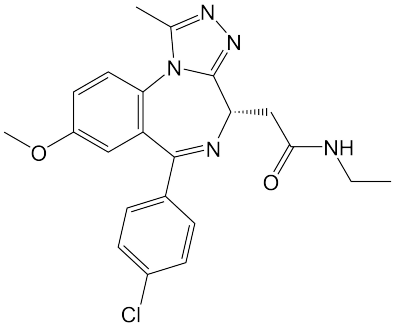
Es gibt Proteine, welche eine Acetylierung der Histone vornehmen, Writer (HATs) oder diese entfernen, Eraser (HDACs), sowie Proteine, die diese Markierungen lesen können (Reader). BRD4 gehört zu den Readern und beeinflusst so zusammen mit dem positiven Transkriptionselongationsfaktor P-TEFβ die Transkription der Targetgene z. B. *MYC* (siehe Abbildung II.0.4). BRD4 ist ein aktuelles Arzneimittelziel (*drug target*) zur Behandlung von Krebs, inflammatorischen Krankheiten, viralen Infektionen und kardiovaskukläreren Erkrankungen (Filippakopoulos und Knapp 2014; Ferri, Petosa und McKenna 2016). In klinischen Studien befinden sich bereits einige Inhibitoren der BET-Proteinfamilie (BETi) zur Behandlung von Krebs und kardiovaskuklären Erkrankungen. In Abbildung II.0.5 sind drei Beispiele von in klinischen Studien befindlichen BETi dargestellt. Eine Übersicht dazu geben unter anderem Ferri, Petosa und McKenna (2016). RVX-208 (siehe Abbildung II.0.5a) ist ein Derivat des Polyphenols Resveratrol und befindet sich in klinischen Studien der Phase III zur Behandlung von Patienten mit Typ II Diabetes und kardiovaskulären Erkrankungen. OTX-015 (siehe Abbildung II.0.5b) und I-BET 762 (siehe Abbildung II.0.5c) sind Derivate der *chemical probe* JQ1 und befinden sich zur Behandlung von verschiedenen Tumoren z. B. NUT-Karzinom in den klinischen Phasen I-II.



(a) RVX-208



(b) OTX-015



(c) I-BET 762

Abbildung II.0.5.: BETi in klinischen Studien: RVX-208 (a), OTX-015 (b) und I-BET 762 (c).

II.2.2. PPARγ

Der Peroxisome proliferator-activator receptor γ (PPARγ) ist ein nukleärer Rezeptor und Transkriptionsfaktor, welcher eine zentrale Rolle in entzündlichen Prozessen sowie allgemeiner unter oxidativem Stress und der Energiehomöostase einnimmt (Polvani, Tarocchi und Galli 2012). Einige mit PPARγ verbundene und für diese Arbeit bedeutende Signalwege sowie ihre gegenseitige Beeinflussung und Auswirkung auf die Transkription sind in Abbildung II.0.6 zusammengefasst. Neben PPARγ gibt es zwei weitere Subtypen PPARα und PPARδ. PPARγ bildet mit dem Retinoid X Rezeptor ein Heterodimer, welches an PPAR Response Elemente (PPREs) in den Promoterregionen von Zielproteinen bindet (Varga, Czimmerer und Nagy 2011). Durch Ligandaktivierung von PPARγ kann die Transkription der Zielproteine beeinflusst werden. PPARγ ist ein Arzneimittelziel für diverse Erkrankungen, wie Diabetes, Krebs und Atherosklerose (Kersten, Desvergne und Wahli 2000). Als Antidiabetika wurden Vertreter der Thiazolidindione (Glitazone) zugelassen, welche PPARγ-Agonisten sind und deren Nutzen-Risiko-Bewertung kontrovers diskutiert wird. Ein natürlicher Ligand ist Isosilybin A siehe Abbildung II.0.7a (Zhang *et al.* 2016). Interessanterweise besitzt Isosilybin A ein ähnliches Grundgerüst wie Kinase- und BRD4-Liganden (Flavopiridol, Abbildung II.0.7b).

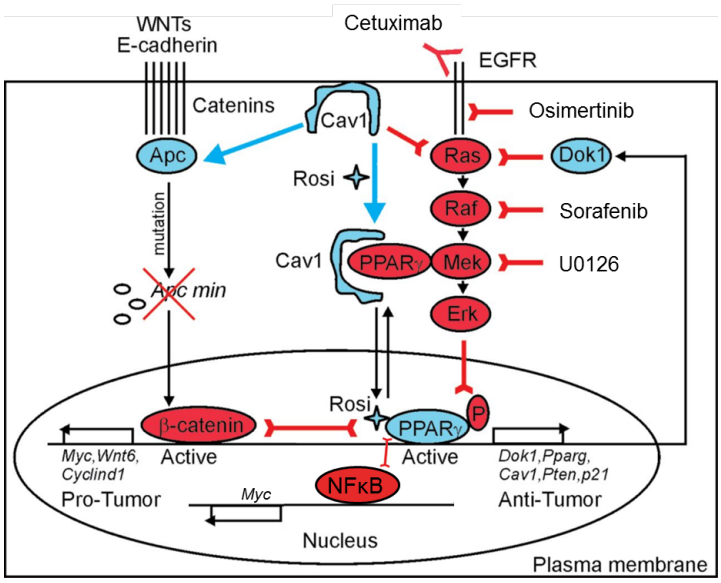
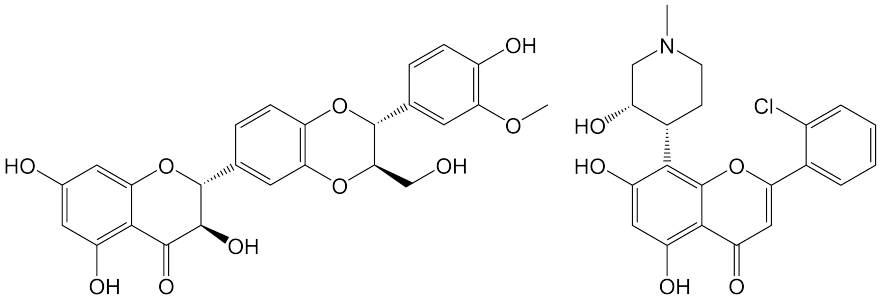


Abbildung II.0.6.: Funktionen von PPARγ und ausgewählte mit PPARγ verbundene Signalwege: Wnt/β-Catenin, RAS (*Rat sarcoma*) und NFκB. Rote Pfeile: inhibitorische Wirkung, blaue Pfeile: aktivierende oder stabilisierende Wirkung, Rosi: Rosiglitazon. Verändert nach Friedrich *et al.* (2013).



(a) Isosilybin A (b) Falvopiridol

Abbildung II.0.7.: Der natürliche PPARγ-Ligand Isosilybin A (a) hat ein ähnliches Grundgerüst wie Falvopiridol, ein Kinase- und BRD4-Inhibitor (b).

II.2.3. Crosstalk von BRD4 und PPARγ

In diesem Abschnitt sollen die physiologischen Funktionen von BRD4 und PPARγ mit Schwerpunkt auf den gemeinsamen Prozessen dargelegt werden. BRD4 und PPARγ besitzen beide eine wichtige Funktion in Tumorerkrankungen (Zhang *et al.* 2013; Aouali *et al.* 2015; Reddy, Lakshmi und Reddy 2016), Adipositas (Zhang *et al.* 2013; Zhang *et al.* 2016; Kersten, Desvergne und Wahli 2000), Diabetes mellitus Typ II (Kersten, Desvergne und Wahli 2000; Filippakopoulos und Knapp 2014; Polvani, Tarocchi und Galli 2012; El Akoum 2014), inflammatorischen Erkrankungen (Zhang *et al.* 2013; Kotla und Rao 2015; Tasdemir *et al.* 2016) besonders Atherosklerose (Kersten, Desvergne und Wahli 2000; Filippakopoulos und Knapp 2014; Kotla und Rao 2015; Brown *et al.* 2014) und kardiovaskulären Erkrankungen (Zhang *et al.* 2016; Filippakopoulos und Knapp 2014; El Akoum 2014; Ferri, Petosa und McKenna 2016; Shi und Vakoc 2014). Auffällig ist, dass alle genannten Krankheiten mit der oxidativen Stressantwort in Verbindung stehen. Unter Einfluss von oxidativem Stress kommt es zu einer Reprogrammierung der Genexpression. Über einen längeren Zeitraum der oxidativen Stresseinwirkung, wie bei chronischen entzündlichen Erkrankungen, kann es zu einer permanenten Umprogrammierung kommen, welche den Verlust von Zelllinien-spezifischen Expressionsmustern zur Folge haben kann und in Tumorgenese resultieren kann. Unter akuter oxidativer Stresseinwirkung ist die Umprogrammierung jedoch vorteilhaft und führt zur Apoptose von infizierten oder seneszenten Zellen. Damit wirkt sie präventiv der Tumorgenese entgegen. Für den Umprogrammierungsprozess muss die Transkriptionsmaschinerie neu verteilt werden. Neue Super Enhancer werden gebildet, um die massive Expression von Stress-assoziierten Genen zu fördern. BRD4 und der Mediatorkomlex (insbesondere MED1) stabilisieren sich gegenseitig. Beide sind ko-lokalisiert an Super Enhancern (Shi und Vakoc 2014) und BRD4 ist ein *enhancer factor* (Brown *et al.* 2014). MED1 ist ein Teil des Mediatorkomplexes und zugleich ein Zielprotein von PPARγ (Hussong *et al.* 2014; Yin und Wang 2014). Zudem interagiert BRD4 mit MED1 (Bhagwat *et al.* 2016). Des Weiteren interagiert MED1 mit PPARγ mithilfe des LXXLL Motivs vieler nukleärer Rezeptoren. Interessanterweise besitzt die BD1 der BRD4 (kurz: BRD4(1)) ebenfalls ein LXXLL Motiv.
NFκB (*nuclear factor 'kappa-light-chain-enhancer' of activated B-cells*) ist ein Schlüssel-Transkriptionsfaktor der inflammatorischen Antwort (Filippakopoulos und Knapp 2014). Bemerkenswert ist, dass BRD4 ein Ko-Aktivator von NFκB ist (Ferri, Petosa und McKenna 2016) und acetyliertes NFκB binden kann (Filippakopoulos und Knapp 2014). Andererseits wird NFκB durch PPARγ inhibiert (siehe Abbildung II.0.6) (Polvani, Tarocchi und Galli 2012). Dies lässt eine Gegenspielerrolle von BRD4 und PPARγ in diesem Prozess vermuten. Ein wichtiges Protein der oxidativen Stressantwort ist NRF2. PPARγ-Agonisten erhöhen die NRF2-Expression (Polvani, Tarocchi und Galli 2012) wohingegen BRD4 die Expression von KEAP1 erhöht und somit die NRF2 Aktivität reduziert (Hussong *et al.* 2014).

Ein weiterer starker Hinweis auf eine funktionelle Beziehung zwischen BRD4 und PPARγ ist, dass BRD4 die Expression von PPARγ2 über JMJD6 regulieren kann (Hu *et al.* 2015). Außerdem ist BRD2 ein Ko-Repressor von PPARγ (Denis, Nikolajczyk und Schnitzler 2010). Während der Adipogenese ist BRD4 ein Regulator der PPARγ-Expression (Brown *et al.* 2018). Weitere Proteine (beziehungsweise deren Expression), die durch BRD4 und PPARγ beeinflusst werden, sind BCL2 (Polvani, Tarocchi und Galli 2012; Shi und Vakoc 2014) und CAT (Polvani, Tarocchi und Galli 2012; Hussong *et al.* 2014).
Die Proteinexpression ist ein strikt regulierter Vorgang, bei dem posttranslationalen Modifikationen (PTMs) von zentraler Bedeutung für die Kontrolle des Prozesses sind. Es ist in diesem Zusammenhang interessant, dass Proteine, welche PTMs schreiben oder lesen, wie Kinasen und BRDs, ähnliche oder sogar gleiche Liganden aufweisen. Weitere Aspekte des Crosstalks zwischen BRD4 und PPARγ insbesondere im Zusammenhang mit Resistenzen werden unter dem Abschnitt III.3.2.10 diskutiert.

II.3. Benchmarkdatensätze

Eine fundamentale Voraussetzung zur Entwicklung von prädiktiven Algorithmen ist eine ausreichende und qualitativ hochwertige Datengrundlage. Einerseits werden Daten zur Entwicklung des Algorithmus und als Eingabe verwendet und andererseits werden sie zur Validierung sowie Evaluierung benötigt. Konkret für diese Arbeit bedeutet dies, dass auf Grundlage von Bioaktivitätsdaten bioähnliche Substrukturen, wie Bioisostere oder pharmakophore Gruppen und Atomtypen, identifiziert werden. Unter diesem Vorwissen kann die Ähnlichkeitsbewertung durch die verwendeten neuen Algorithmen auf die Erfassung von Bioähnlichkeit zugeschnitten werden. Ob diese Einbeziehung der gefundenen Bioähnlichkeiten die Performanz des Verfahrens steigert, d. h. bioähliche Moleküle angereichert werden können (innerhalb der vorderen Ergebnisse einer Ähnlichkeitssuche oder innerhalb eines Clusters), wird anschließend mithilfe von weiteren Bioaktivitätsdaten evaluiert. Bei den in dieser Arbeit verwendeten Algorithmen handelt es sich insofern um prädiktive Algorithmen, als nach dem SPP von einer molekularen Ähnlichkeit oft auf eine Bioähnlichkeit geschlossen werden kann. Da in dieser Arbeit biologische Daten berücksichtigt werden und biologische Systeme aufgrund ihrer Komplexität eine inhärent große Streubreite aufweisen, stellt die Verfügbarkeit einer geeigneten Datengrundlage eine Herausforderung dar. In den folgenden Abschnitten werden zwei zur Analyse der neu entwickelten MSS-Methode verwendete Datensätze und ein interessanter, jedoch kritisch zu beurteilender Datensatz beschrieben.

II.3.1. Benchmarkdatensatz von Stahl *et al.* (SBDS)

Der Datensatz von Stahl *et al.* (2005) besteht aus den Liganden von neun Proteinen: Acetylcholinesterase (AChE), Angiotensin-II-Rezeptor, Cyclin-abhängige Kinase 2 (CDK2), Cyclooxygenase-2 (COX-2), Dopamin D4-Rezeptor, Estrogenrezeptor (ER), Histamin-H3-Rezeptor, Matrix-Metalloproteinase 3 und Thrombin. Er umfasst insgesamt 466 Moleküle. Eine Aufteilung der Moleküle des Datensatzes auf die Target-Proteine ist in Tabelle II.0.1 zu finden. Beim SBDS handelt es sich um einen bereits vorgeclusterten Datensatz, bei dem aus insgesamt 653 Molekülen nur die Singletons und Clusterzentren ausgewählt wurden (Stahl *et al.* 2005). So wurde ein anspruchsvoller Datensatz für Clustering-Algorithmen erhalten, der keine trivialen Ähnlichkeiten, wie sie z. B. bei Substanzserien vorkommen, enthält. Dieser Datensatz wurde zur Evaluierung des StruClus-Algorithmus verwendet (siehe Abschnitte III.1.2.2 und III.2.2.5).

Target-Protein	Anzahl Liganden
AChE	40
D4	68
An2	55
ER	44
MMP	75
H3	23
Thrombin	37
COX-2	79
CDK2	45

Tabelle II.0.1.: Zusammensetzung des SBDS nach Target-Proteinen. Proteinabkürzungen: Acetylcholinesterase (AChE), Angiotensin-II-Rezeptor (An2), Cyclin-abhängige Kinase 2 (CDK2), Cyclooxygenase-2 (COX-2), Dopamin D4-Rezeptor (D4), Estrogenrezeptor (ER), Histamin-H3-Rezeptor (H3), Matrix-Metalloproteinase 3 (MMP).

II.3.2. Benchmarkdatensatz von O'Boyle und Sayle (OBDS)

O'Boyle und Sayle (2016) entwarfen einen Datensatz basierend auf ChEMBL Daten (Bento *et al.* 2014) zum Vergleich von Fingerprints. In dieser Arbeit wurde der OBDS zur Evaluierung des BBPMCS-Algorithmus verwendet. Der Aufbau des OBDS ist schematisch in der Abbildung II.0.8 dargestellt. Der OBDS besteht aus zwei Subsets zur Evaluierung der Erkennung von einerseits sehr ähnlichen und andererseits weiter entfernt ähnlichen Molekülen. In der Publikation von O'Boyle und Sayle (2016) werden zwei Moleküle als ähnlich angesehen, wenn es wahrscheinlich wäre, dass sie ein medizinischer Chemiker zur selben Zeit innerhalb desselben Projektes synthetisieren und testen würde.

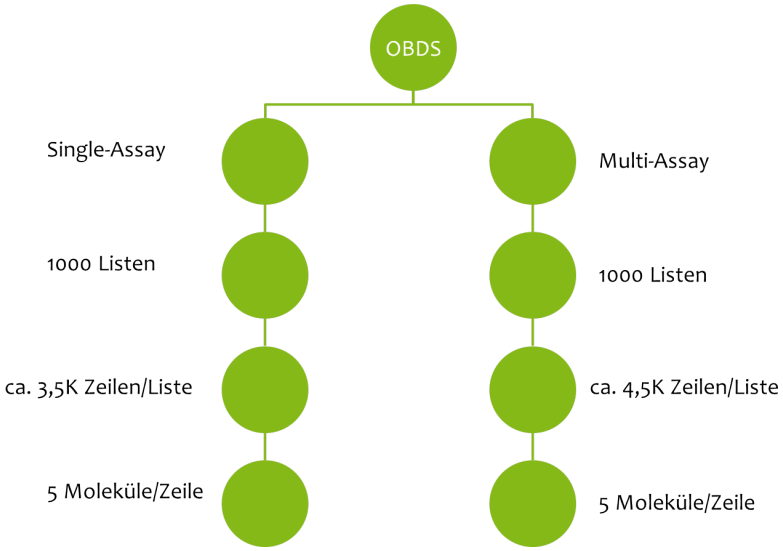


Abbildung II.0.8.: Aufbau des O'Boyle und Sayle Datensatzes (OBDS).

Für das erste Subset, Single-Assay genannt, werden für einen Vergleich nur die Moleküle innerhalb einer Publikation betrachtet, wohingegen für das zweite Subset, Multi-Assay genannt, Moleküle eines Vergleiches mehreren Publikationen entstammen, welche jedoch immer mindestens ein Molekül gemeinsam haben müssen. Jedes dieser Subsets besteht wiederum aus mehreren Listen von 5-Tupel (Quintupel). Die Quintupel sind nach absteigender Ähnlichkeit zum ersten Molekül, der Referenz, sortiert. Die Referenz wird gegen die anderen vier Moleküle des Quintupels verglichen. Eine gute Methode zum Vergleich der Ähnlichkeiten von niedermolekularen Verbindungen sollte dementsprechend bei der Suche mithilfe der Referenz die vier anderen Moleküle entsprechend absteigend ranken. Eine Unabhängigkeit von einzelnen Testdaten soll durch die tausendfache Ausführung von mindestens dreitausend Testreihen gewährleistet werden. Je häufiger eine Methode die Ähnlichkeit richtig einstuft, desto besser ist sie und desto höher wird sie im Ergebnisgraph einsortiert. Die Zahlen an den Kanten geben die Differenzen der Anzahl der Datensätze an, welche von den zwei Methoden an den jeweiligen Enden der Kanten besser bezüglich der vorgegebenen Ähnlichkeit eingestuft wurden.

II.3.3. Benchmarkdatensatz von Backmann *et al.* (BBDS)

Dieser Datensatz befasst sich mit dem Problem, dass inaktive Verbindungen selten publiziert werden (*Publikationsbias*). Aufgrund der immensen Bedeutung einer validen Datengrundlage soll dieser Datensatz beispielhaft vorgestellt werden, auch wenn er nicht für Experimente in der vorliegenden Arbeit verwendet wurde. Der Entwicklungsprozess und die solide Evaluierung eines prädiktiven Algorithmus sind auf eine Negativkontrolle (hier inaktive Verbindungen) angewiesen.

Leider liegen jedoch nur wenige Daten zu inaktiven Verbindungen in öffentlich zugänglichen Datenbanken vor. Zudem variiert die Definition von „inaktiv“ (keine messbare Aktivität bei einer Konzentration von 50 μM, 10 μM, IC_{50} größer 20 μM). Eine Datenbank, welche inaktive Verbindungen zusammen mit der Definition von „inaktiv“ beinhaltet, ist PubChem Bioassay (Wang *et al.* 2017). Diese dient als Grundlage für den Datensatz von Backmann *et al.* (Backman, Evans und Girke 2017). Der BBDS stellt eine Matrix von circa 65.000 Verbindungen und 260 Proteinen zur Verfügung. Eine solche Matrix wäre sehr wertvoll, doch dazu müsste eine zuverlässige Annotation von aktiven, inaktiven und nicht getesteten Verbindungen erfolgen. In der Arbeit von Backman, Evans und Girke (2017) werden allerdings inaktive und nicht getestete Verbindungen gleich behandelt. Eine Unterscheidung ist jedoch essentiell, da über die nicht getesteten Verbindungen keine Aussage getroffen werden kann. Bewertet also ein Algorithmus eine nicht getestete Verbindung hoch im Bezug auf eine aktive Verbindung, sollte dies nicht zur Annahme einer fehlerhaften Einschätzung durch den Algorithmus führen (im Gegensatz zu einer hohen Bewertung einer inaktiven Verbindung). Genauso sollte ein Algorithmus nicht daraufhin optimiert werden, nicht getestete Verbindungen schlecht zu bewerten, wie es für inaktive Verbindungen plausibel wäre.

Teil III.

Ergebnisse und Diskussion

Das komplexe Thema der Ähnlichkeit von niedermolekularen Verbindungen, wurde in Teilprojekte untergliedert und dargestellt, um es aus mehreren Gesichtspunkten bzw. auf verschiedenen Ebenen beleuchten und untersuchen zu können. Jedes Teilprojekt wird mit einer prägnanten Einleitung eingeführt. Anschließend werden die Ergebnisse präsentiert und diskutiert, abgeschlossen wird mit einer Zusammenfassung. Im Kapitel III.1 wird die Analyse und Erstellung einer innovativen Datenbank, genannt CH*I*PMUNK, im Detail beschrieben. Diese innovative und leistungsstarke Datenbank ist für die Anwendung der Ergebnisse in den darauffolgenden Kapiteln von Bedeutung. In ihr können mithilfe der Ergebnisse aus den anderen Teilprojekten in Bezug auf die Bioaktivität ähnliche Moleküle gefunden werden, die noch nicht synthetisiert wurden und die den chemischen Raum erweitern (teilweise auch außerhalb der Lipinski-Regeln), jedoch mit hoher Wahrscheinlichkeit synthetisierbar sind. Die Ähnlichkeit von niedermolekularen Verbindungen wurde auf verschiedenen Ebenen untersucht. Im Kapitel III.2 werden die Ebenen der Ähnlichkeit von niedermolekularen Verbindungen basierend auf Atomtypengruppen und pharmakophoren Gruppen untersucht. Anschließend folgt im Kapitel III.3 die Analyse der Scaffold-basierten Ähnlichkeit. In die Untersuchung der Scaffold-Ebene fließen weitere Informationen zu den Proteinen ein (z. B. deren räumliche Struktur) und ergeben die Protein-Ebene. Dies ist besonders wichtig, da die Ähnlichkeit von niedermolekularen Verbindungen in der vorliegenden Arbeit aus der Sicht der Proteine untersucht werden soll, um so eine Ähnlichkeit bezüglich der Bioaktivität basierend auf der molekularen Ähnlichkeit ableiten zu können. Des Weiteren wurden die auf der Scaffold- und Protein-Ebene gefundenen Ähnlichkeiten am Beispiel neuer Inhibitoren für BRD4, einer sehr aktuellen hoffnungsvollen Zielstruktur für neue Arzneistoffe zur Therapie von z. B. Tumoren oder kardiovaskulären Erkrankungen, experimentell überprüft. Abschließend erfolgen Einblicke in die Ebenen der bioisosteren Ähnlichkeit sowie der Form (Shape) basierten Ähnlichkeit in Kapitel III.4.

III.1. Erstellung und Analyse von CH*/*PMUNK

III.1.1. Einleitung

Wie in Humbeck und Koch (2017) dargestellt, existieren bereits mehrere Arten von Molekül-Datenbanken. Dazu gehören Datenbanken synthetisierter Moleküle wie z. B. die ZINC (Irwin und Shoichet 2005; Sterling und Irwin 2015) oder Datenbanken kommerziell erhältlicher Verbindungen von den jeweiligen Anbietern wie MolPort (*MolPort* 2018) oder eMolecules (*eMolecules* 2018). Des Weiteren gibt es Datenbanken mit Bioaktivitätsdaten getesteter Verbindungen, wie ChEMBL (Bento *et al.* 2014; Gaulton *et al.* 2017), PubChem BioAssay (Wang *et al.* 2017) oder DrugBank (Law *et al.* 2013) und Datenbanken von virtuellen Molekülen, wie die GDB-17 (Ruddigkeit *et al.* 2012) und SCUBIDOO (Chevillard und Kolb 2015). Jede dieser Datenbanken ist aufgrund ihres speziellen Anwendungsbereichs wichtig und wertvoll. So geben die ersten beiden genannten Datenbankarten einen Überblick über den synthetisierten chemischen Raum, während Datenbanken mit Bioaktivitätsdaten ebenfalls einen Einblick in den bioaktiven Raum gewähren. Virtuelle Datenbanken sind ein wertvolles Werkzeug zur Erweiterung des bekannten chemischen Raums. Sie besitzen das Potential fundamental neuartige Verbindungen aufzuzeigen.
Unter dem in dieser Arbeit auf verschiedenen Ebenen behandelten Aspekt der Ähnlichkeit niedermolekularer Verbindungen können solche Datenbanken verwendet werden, um ausgehend von einer bekannten, bioaktiven Verbindung, innovative Vorschläge von chemisch verschiedenen, aber dennoch bioähnlichen Substanzen zu generieren. Eine weitere mögliche Anwendung besteht in der Unterstützung bei der Targetidentifizierung in phänotypischen Screens. Ein phänotypischer Hit kann gegen eine Datenbank von bioaktiven Molekülen mit bekannter Zielstruktur (Target) abgeglichen werden und so auf potentielle Zielstrukturen des Hits geschlossen werden. Zur Umsetzung beider Anwendungen wird eine Datenbank benötigt, die zum einen einen neuen chemischen Raum abdeckt und zum anderen Bioaktivitätsdaten beinhaltet. Außerdem besteht für den neuen chemischen Raum die Notwendigkeit der einfachen Realisierbarkeit im Labor. Dieses Kriterium schließt die Verwendung der GDB-17 aus, da zur Erstellung nur grundlegende chemische Anforderungen, wie Valenzen, berücksichtigt wurden. SCUBIDOO basiert hingegen auf Reaktionen, die häufig in der medizinischen Chemie verwendet werden (Chevillard und Kolb 2015). Jedoch beinhaltet SCUBIDOO keine Bioaktivitätsdaten.

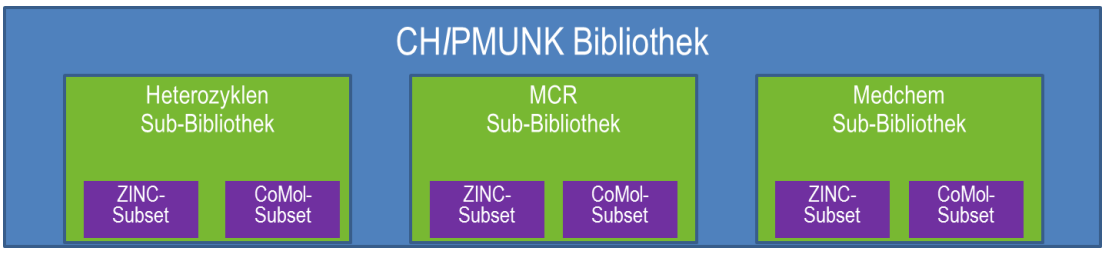


Abbildung III.1.1.: Übersicht über den Aufbau von CH*I*PMUNK.

Um den chemischen Raum zu erweitern und damit eine höhere Diversität zu erzielen, wurde eine neue Datenbank erstellt, welche aus drei Arten von Reaktionen zusammengesetzt ist (siehe Abbildung III.1.1): (i) häufig verwendeten Reaktionen in der medizinischen Chemie nach Hartenfeller *et al.* (2011) (medchem), (ii) Multikomponentenreaktionen nach Koes, Dömling und Camacho (2018) und *Anchor-Biased MCR Reactions* (2018) (MCR) und (iii) Reaktionen aus SciFinder (*SciFinder* 2018), welche die typischen Heterozyklen in Arzneistoffen darstellen nach Taylor, MacCoss und Lawson (2014) (heterozyklen). Damit enthält die neue Datenbank zwei weitere Reaktionsarten gegenüber SCUBIDOO. Die Reaktionen wurden jeweils mit zwei Sets von Edukten *in silico* durchgeführt (siehe auch Abschnitte IV.2.5 und IV.2.6). Diese stammten entweder (i) aus der ZINC-Datenbank (ZINC) oder (ii) von eMolecules und MolPort (CoMol). Darüberhinaus wurde die erstellte Datenbank, genannt CH*I*PMUNK (CHemically feasible *In silico* Public Molecular UNiverse Knowledge base), mit der ZINC, der CoMol und der ChEMBL vereinigt, um einen Vergleich zum synthetisierten und als Edukt verwendeten Raum sowie zum bioaktiven Raum zu ermöglichen. In diesem Kapitel werden die Eigenschaften der neu erstellten CH*I*PMUNK beleuchtet und ihre Tauglichkeit im Hinblick auf die Entdeckung neuer Arzneistoffe untersucht. Der Aufbau von CH*I*PMUNK kann der Abbildung III.1.1 entnommen werden.

III.1.2. Ergebnisse und Diskussion

Insgesamt umfasst CH*I*PMUNK mehr als 95 Millionen Verbindungen. Eine Übersicht über den Prozentsatz an Verbindungen, welche die Lipinski-Regel erfüllen (*rule of five compliant* (innerhalb der Lipinski-Regel) (Ro5)) (Lipinski *et al.* 1997) sowie den Prozentsatz an Verbindungen, welche einen *synthetic accessibility* (synthetische Zugänglichkeit) (SA) Score (Ertl und Schuffenhauer 2009) größer Null besitzen und die Anzahl an Produkten je Subset ist in Tabelle III.1.1 für die Produkte basierend auf CoMol-Edukten und in Tabelle III.1.2 für die Produkte basierend auf ZINC-Edukten zu finden. Ebenfalls untersucht wurden diese Eigenschaften für jenen Anteil jedes Subsets, welcher nur Verbindungen mit einer Molekülmasse von kleiner 500 Da beinhaltet (im Folgenden 500-Anteil genannt).

Der SA-Score ist ein Versuch die synthetische Zugänglichkeit einer Verbindung in einer berechenbaren Zahl auszudrücken. Dazu werden unter anderem die Molekülmasse sowie die Komplexität des Moleküls und Fragment-basierte Beiträge in die Kalkulation miteinbezogen (Ertl und Schuffenhauer 2009). In der verwendeten Implementierung in Pipeline Pilot (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015) sind höhere Werte des SA-Scores ein Indiz für eine leichtere synthetische Zugänglichkeit des Moleküls. Die tatsächliche Zugänglichkeit sowie zum Beispiel die Ausbeute lassen sich jedoch nur im Labor überprüfen bzw. bestimmen. Allerdings dient der SA-Score zur ersten präexperimentellen tendenziellen Einschätzung der Synthetisierbarkeit. Die medchem Sub-Bibliothek zeigt einen stabilen hohen Anteil an Molekülen mit einem SA-Score größer Null von 48,9 % bzw. 58,7 % für das Subset basierend auf CoMol- respektive basierend auf ZINC-Edukten. Sowohl die medchem als auch die heterozyklen Sub-Bibliothek weist, insbesondere für den ungefilterten Datensatz, einen deutlichen Anteil (75,1 % und 25,9 % für CoMol bzw. 14,3 % und 59,7 % für ZINC) an Ro5-Molekülen auf. Für die MCR Sub-Bibliothek ist ein geringerer Anteil an Ro5-Verbindungen und ebenfalls ein geringerer Anteil an Verbindungen, welche einen SA-Score größer Null besitzen, im Vergleich zu den beiden anderen Sub-Bibliotheken, zu erkennen. Dies liegt darin begründet, dass beide Deskriptoren von der Molekülmasse abhängig sind. Deutlich wird dies durch die Betrachtung der Eigenschaften des 500-Anteils. Da bei einer MCR mehr als zwei Edukte unter hoher Atomeffizienz, d. h. ein hoher Anteil der Atome der Edukte findet sich im Produkt wieder, reagieren, ist die Verteilung der Molekülmasse zu höheren Massen verschoben (siehe Abbildungen III.1.2 M und III.1.4 M). Nichtsdestotrotz sind MCR mit hoher Effizienz durchführbar (Dömling, Wang und Wang 2012) und somit häufig gut synthetisch zugänglich sowie eine interessante Alternative zur Generierung von Protein-Protein-Interaktions-Inhibitoren (PPIIs). PPIIs besitzen im Vergleich zu Protein-Ligand-Inhibitoren eine größere Molekülmasse, einen höheren logP-Wert, eine größere *polar surface area* (polare Oberfläche) (PSA) sowie eine größere Anzahl an Wasserstoffbrückenakzeptoren und rotierbaren Bindungen. Außerdem besitzen sie eine größere Anzahl an Mehrfachbindungen (entsprechend einer größeren Anzahl an Doppelbindungen) und eine größere Anzahl an Atomen in aromatischen Ringen (entsprechend einer größeren Anzahl an aromatischen Bindungen) (Sperandio *et al.* 2010).

Im Vergleich zu den Edukt-Datenbanken (CoMol und ZINC) besitzen die Subsets von CH*I*PMUNK einen geringeren Anteil an Ro5-Molekülen. Abgesehen davon, dass die Lipinski-Regel eher einen groben Richtwert als eine harte Grenze für die orale Verfügbarkeit von niedermolekularen Verbindungen darstellt, weckt zunehmend der chemische Raum, der über den von den Lipinski-Regeln aufgespannten Raum hinausgeht (*beyond rule of five* (außerhalb der Lipinski-Regel) (bRo5)), Interesse (Zhang und Wilkinson 2007). Insbesondere bisher als schwer durch niedermolekulare Verbindungen beeinflussbar geltende Proteine könnten durch den bRo5-Raum für eine therapeutische Beeinflussung zugänglich gemacht werden.

Ein Beispiel für bRo5-Arzneistoffe sind antivirale Proteaseinhibitoren wie der HCV-NS4A-Proteaseinhibitor Daclatasvir mit einer Molekülmasse von 738,88 Da zur Behandlung von Hepatitis C (DeGoey *et al.* 2017). Neben Proteasen und Kinasen sind Transferasen und Isomerasen Ziele von bRo5 klinischen Kandidaten und Arzneistoffen (Doak *et al.* 2016). Bei klassischen Arzneistoffzielen wie Ionenkanälen oder nuklären Rezeptoren sind bRo5-Arzneistoffe allerdings unterrepräsentiert (Doak *et al.* 2016).

	MCR	MCR_{500}	medchem	$medchem_{500}$	heterozyklen	$heterozyklen_{500}$	CoMol
Anzahl Produkte	20.077.099	490.504	36.188.769	8.862.527	10.604.291	4.158.909	9.275.764
% Ro5	2,0	82,2	18,4	75,1	25,9	66,1	83,0
% SA-Score > 0	6,1	53,3	48,9	73,4	63,9	82,5	87,6

Tabelle III.1.1.: Eigenschaften von CH*I*PMUNK basierend auf CoMol-Edukten, der Index 500 kennzeichnet den Anteil eines Subsets, dessen Moleküle eine Molekülmasse von kleiner 500 Da besitzen.

	MCR	MCR_{500}	medchem	$medchem_{500}$	heterozyklen	$heterozyklen_{500}$	ZINC
Anzahl Produkte	11.485.001	300.350	18.279.440	3.634.188	1.088.610	846.078	1.674.502
% Ro5	1,9	71,2	14,3	72,0	59,7	76,8	95,3
% SA-Score > 0	22,8	79,3	58,7	78,6	25,7	32,1	68,1

Tabelle III.1.2.: Eigenschaften von CH*I*PMUNK basierend auf ZINC-Edukten, der Index 500 kennzeichnet den Anteil eines Subsets, dessen Moleküle eine Molekülmasse von kleiner 500 Da besitzen.

III.1.2.1. Physikochemische Eigenschaften von CH*I*PMUNK

Der Deskriptor-Raum von CH*I*PMUNK wurde durch die Berechnung von Histogrammen der Deskriptoren: (i) Molekülmasse (M), (ii) AlogP (berechneter Octanol-Wasser-Verteilungskoeffizient), (iii) Anzahl an Wasserstoffbrückenakzeptoren (HBA), (iv) Anzahl an Wasserstoffbrückendonatoren (HBD), (v) Anzahl an Atomen (N Atome), (vi) Anzahl an Ringen (N Ringe), (vii) Anzahl an aromatischen Ringen (N aromatische Ringe), (viii) Anzahl an rotierbaren Bindungen (N rotierbare Bindungen), (ix) Anzahl an Doppelbindungen (N Doppelbindungen), (x) PSA und (xi) SA-Score für die Subsets beschrieben (siehe Abbildungen III.1.2 und III.1.4). Für den 500-Anteil der Subsets wurden ebenfalls entsprechende Histogramme berechnet (siehe Abbildungen III.1.3 und III.1.5). Aus den Histogrammen ist erkennbar, dass der Deskriptor-Raum der MCR Sub-Bibliothek sich deutlich von den beiden anderen Sub-Bibliotheken (medchem und heterozyklen) unterscheidet.

Da einige Deskriptoren untereinander z. B. mit der Molekülmasse korrelieren, ist eine Erklärung für den distinkten Deskriptor-Raum der MCR Sub-Bibliothek die hohe durchschnittliche Molekülmasse. Interessanterweise sind Unterschiede des Deskriptor-Raumes für die Sub-Bibliotheken deutlicher bei den ZINC-Subsets zu erkennen. Der Deskriptor-Raum des heterozyklen ZINC-Subsets ähnelt am meisten dem Deskriptor-Raum von klassischen Protein-Ligand-Inhibitoren (Ro5) (siehe Abbildung III.1.4 blaue Balken). Das medchem ZINC-Subset weist einen zu größeren Molekülmassen verschobenen Deskriptor-Raum auf, welcher sowohl den Raum von Protein-Ligand- als auch von Protein-Protein-Interaktions-Inhibitoren überspannt (siehe Abbildung III.1.4 grüne Balken). Dieser Datensatz deckt somit voraussichtlich gut den *extended rule of five* (erweiterte Lipinski-Regel) (eRo5) Raum nach Doak *et al.* (2016) ab. Eine weitere Verschiebung des Deskriptor-Raumes zu größeren Molekülmassen ist für das MCR ZINC-Subset erkennbar (siehe Abbildung III.1.4 rote Balken). Folglich ist der Deskriptor-Raum der MCR Sub-Bibliothek dem von PPIIs ähnlich und enthält einen beträchtlichen Anteil an bRo5 Verbindungen. Jedoch befinden sich weiterhin Verbindungen des Ro5-Raumes in dieser Sub-Bibliothek. Der Deskriptor-Raum der heterozyklen und mechchem CoMol-Subsets ist stark überlappend und die in den ZINC-Subsets erkennbaren Differenzen nur marginal ausgeprägt. Das MCR-CoMol-Subset hebt sich erneut durch eine Verschiebung des Deskriptor-Raumes zu größeren Molekülmassen ab. Es fällt auf, dass sich das Bild für die 500-Anteile von CoMol- und ZINC-Subsets umkehrt. Dort sind Unterschiede der Deskriptor-Räume der Sub-Bibliotheken bei CoMol erkennbar (Abbildung III.1.3), während sie für die ZINC-Subsets (Abbildung III.1.5) von untergeordneter Bedeutung sind. Diese Beobachtung stimmt mit den Unterschieden in dem Ro5-Anteil überein. Für den vollständigen Datensatz sind die Unterschiede im Ro5-Anteil in den ZINC-Subsets größer (siehe Tabelle III.1.2) und für den 500-Anteil sind sie in den CoMol-Subsets stärker ausgeprägt (siehe Tabelle III.1.1). Zusammenfassend ist festzustellen, dass CH*I*PMUNK mit seinen drei Sub-Bibliotheken sowohl den Deskriptor-Raum von Protein-Ligand- und Protein-Protein-Interaktions-Inhibitoren als auch von Liganden für bisher als schwer durch (niedermolekulare) Verbindungen zugänglich geltende (undruggable) Proteine (bRo5) umfasst.

III.1.2.2. Analyse des chemischen Raums von CH*I*PMUNK und Vergleich zum bioaktiven Raum

Nach der Analyse des Deskriptor-Raumes erfolgte der Vergleich des chemischen Raumes von CH*I*PMUNK mit dem der Edukt-Datenbanken (ZINC und CoMol). Diese werden als Surrogat-Parameter für den synthetisierten chemischen Raum verwendet. Daneben wurde mit der ChEMBL (Bento *et al.* 2014) eine Bioaktivitätsdatenbank mit in die Analyse einbezogen, um zusätzlich einen Vergleich mit dem bioaktiven Raum zu erlauben. Somit kann einerseits die Abdeckung und die eventuelle Erweiterung des bekannten chemischen Raumes analysiert werden und zum anderen können mögliche Zielstrukturen für neue Verbindungen aufgezeigt werden.

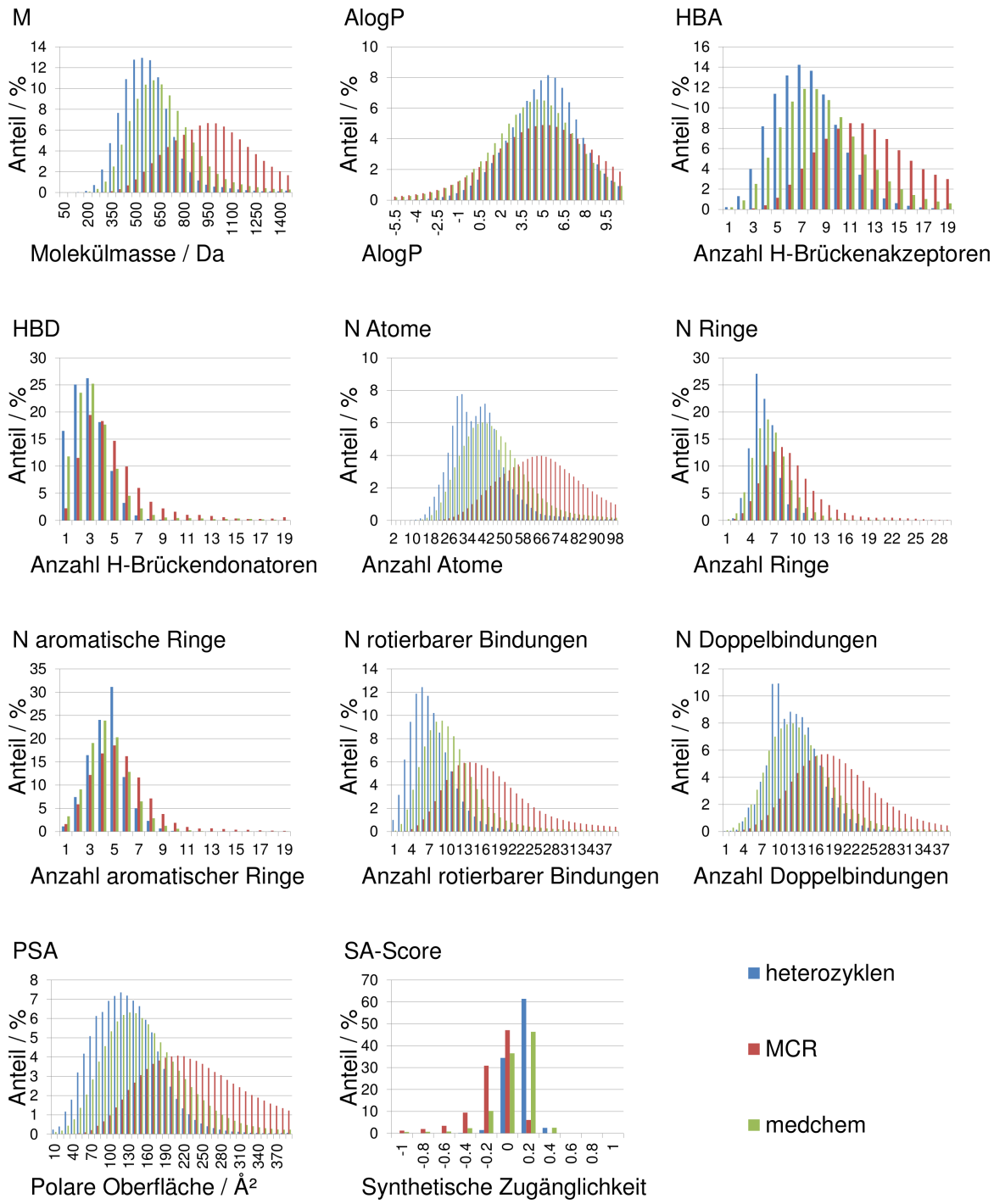


Abbildung III.1.2.: Eigenschaftshistogramme der CoMol-Subsets von CH*I*PMUNK. M: Molekülmasse, AlogP: berechneter Octanol-Wasser-Verteilungskoeffizient, HBA: Anzahl an Wasserstoffbrückenakzeptoren, HBD: Anzahl an Wasserstoffbrückendonatoren, PSA: polare Oberfläche, SA: synthetische Zugänglichkeit und N: Anzahl.

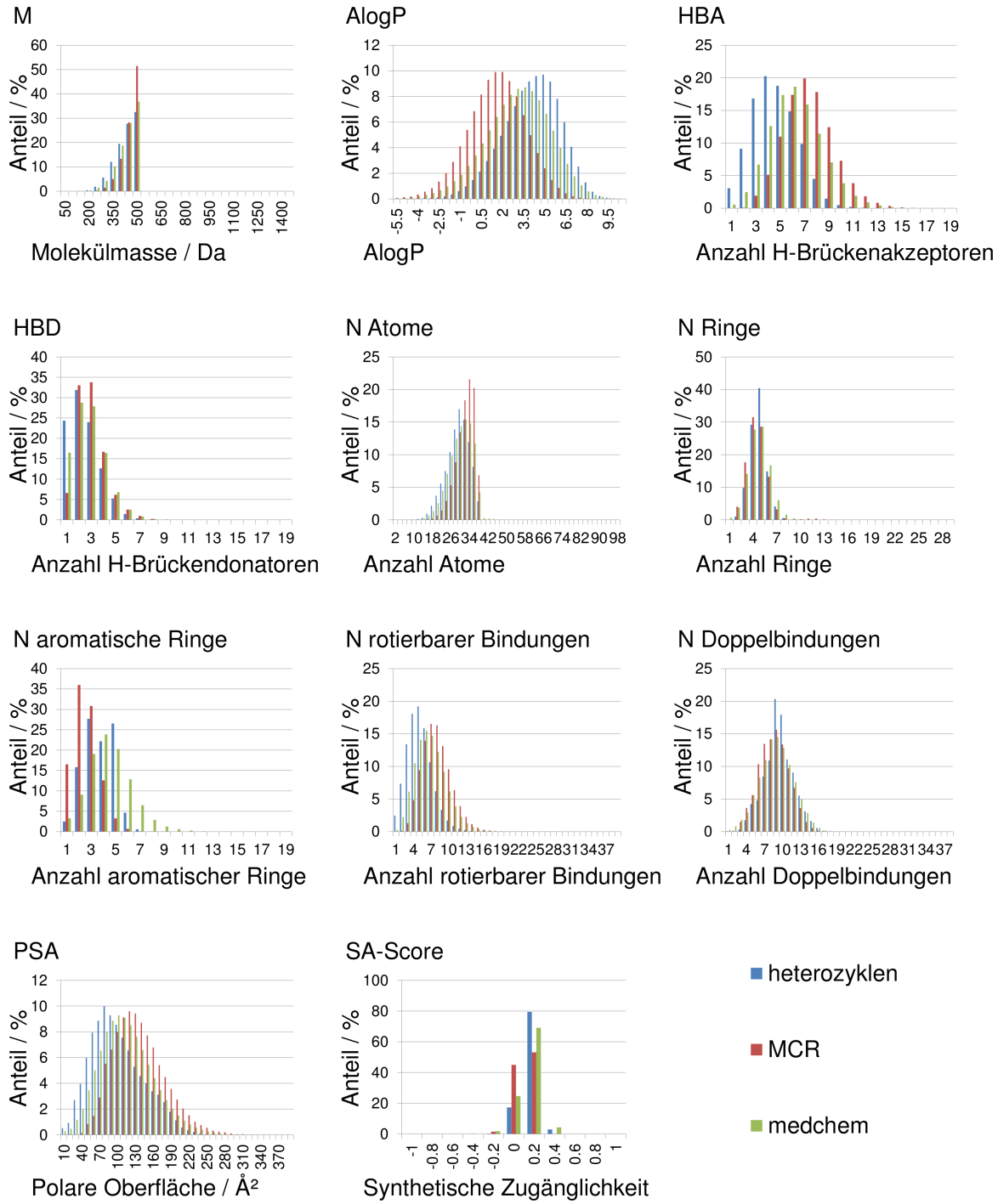


Abbildung III.1.3.: Eigenschaftshistogramme der $CoMol_{500}$-Subsets von CH*I*PMUNK. M: Molekülmasse, AlogP: berechneter Octanol-Wasser-Verteilungskoeffizient, HBA: Anzahl an Wasserstoffbrückenakzeptoren, HBD: Anzahl an Wasserstoffbrückendonatoren, PSA: polare Oberfläche, SA: synthetische Zugänglichkeit und N: Anzahl.

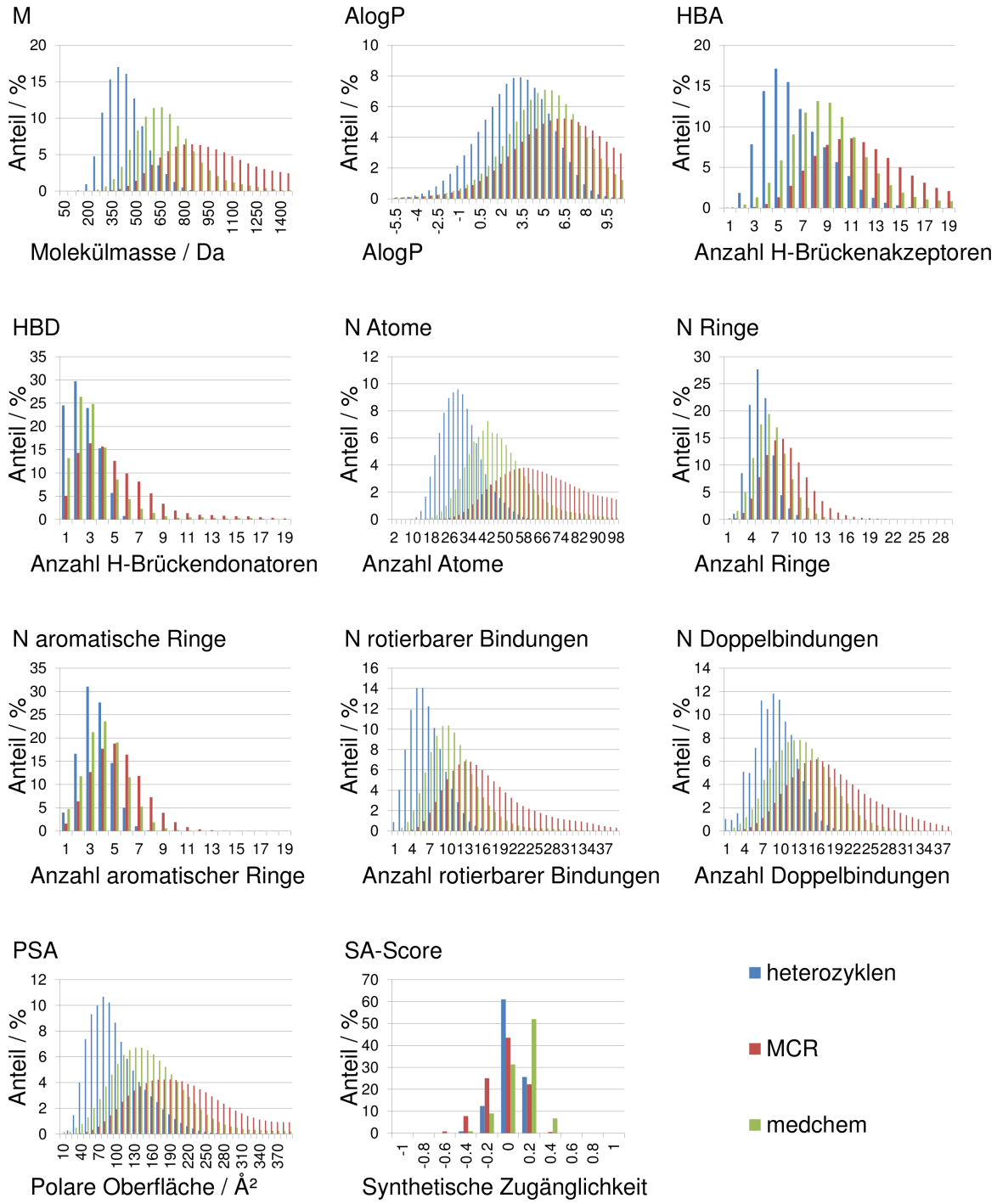


Abbildung III.1.4.: Eigenschaftshistogramme der ZINC-Subsets von CH*I*PMUNK. M: Molekülmasse, AlogP: berechneter Octanol-Wasser-Verteilungskoeffizient, HBA: Anzahl an Wasserstoffbrückenakzeptoren, HBD: Anzahl an Wasserstoffbrückendonatoren, PSA: polare Oberfläche, SA: synthetische Zugänglichkeit und N: Anzahl.

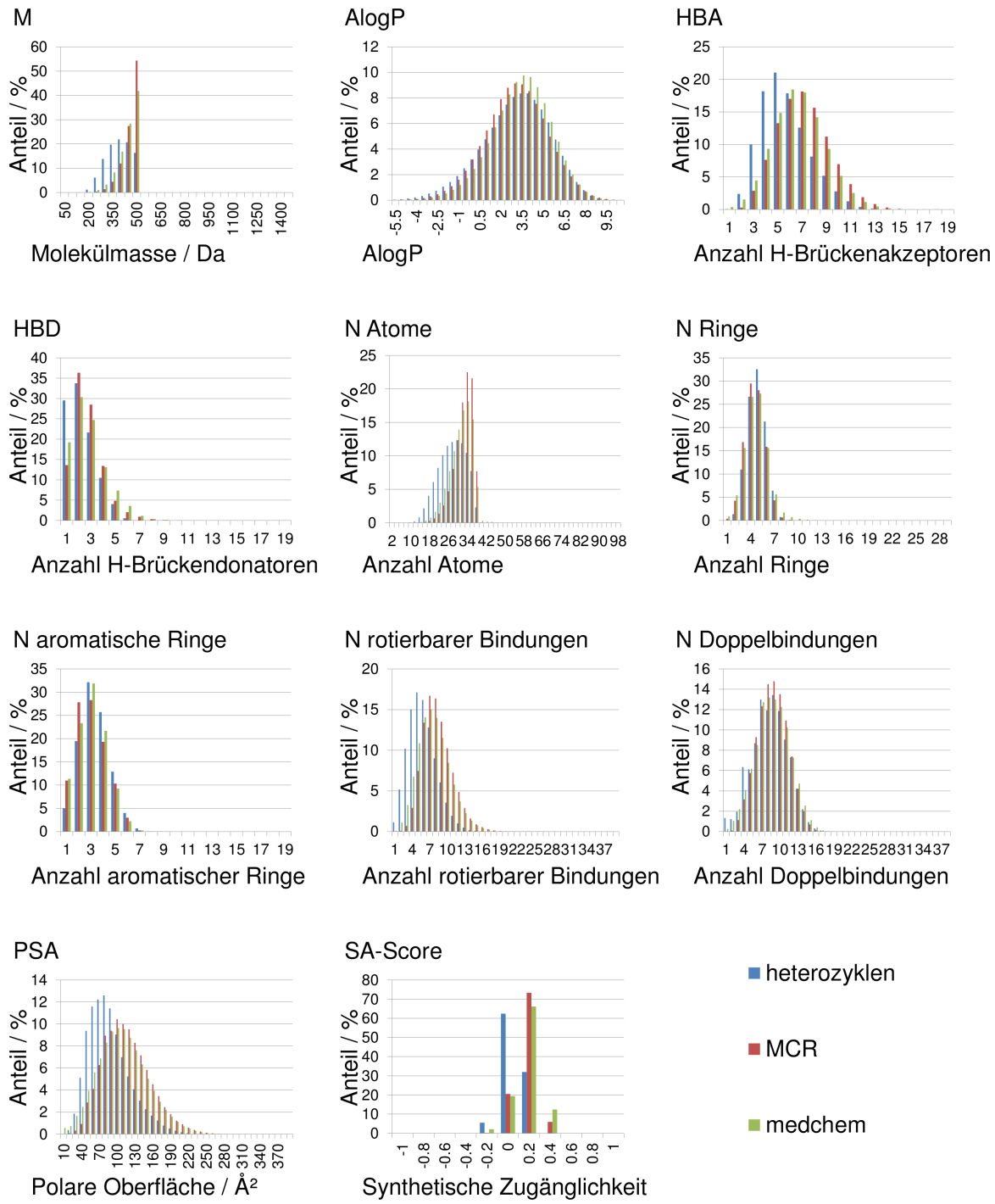
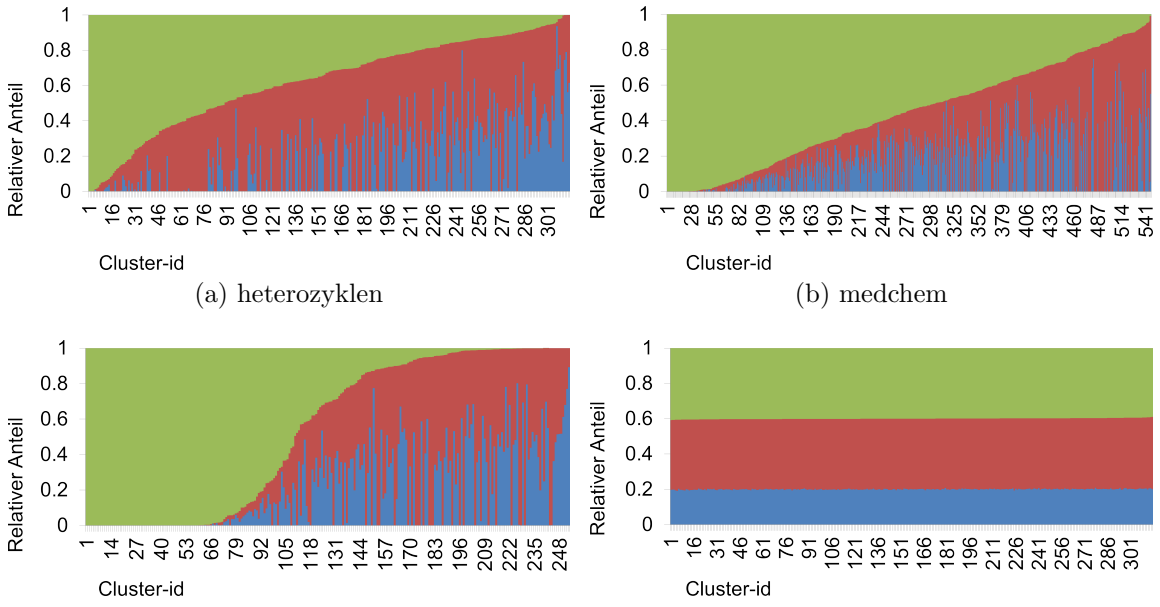


Abbildung III.1.5.: Eigenschaftshistogramme der $ZINC_{500}$-Subsets von CH*I*PMUNK. M: Molekülmasse, AlogP: berechneter Octanol-Wasser-Verteilungskoeffizient, HBA: Anzahl an Wasserstoffbrückenakzeptoren, HBD: Anzahl an Wasserstoffbrückendonatoren, PSA: polare Oberfläche, SA: synthetische Zugänglichkeit und N: Anzahl.



(a) heterozyklen (b) medchem

(c) MCR (d) randomisiertes Clustering

Abbildung III.1.6.: Analyse des chemischen Raumes von CH*I*PMUNK Sub-Bibliotheken: heterozyklen (a), medchem (b), MCR (c) und eine zufällige Verteilung der Verbindungen auf die Cluster (d). Auf der Abszisse sind die Cluster, sortiert nach relativem Anteil an CH*I*PMUNK-Verbindungen und an den Edukt-Datenbanken, aufgetragen. Auf der Ordinate ist der relative Anteil der Verbindungen aus CH*I*PMUNK Sub-Bibliotheken (grün), den Edukt-Datenbanken (CoMol und ZINC) (rot) und der ChEMBL (blau) aufgetragen.

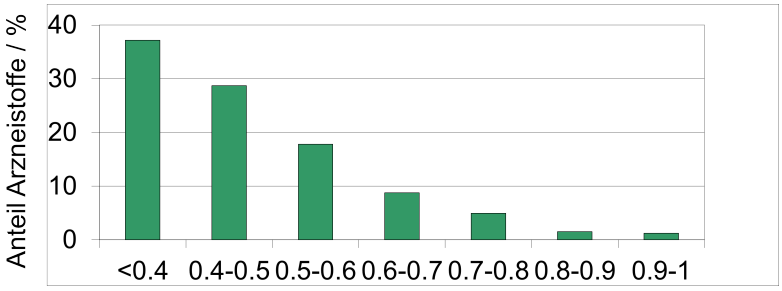
Folglich wurden die drei Sub-Bibliotheken (medchem, heterozyklen und MCR) mit den Edukt-Datenbanken (ZINC und CoMol) sowie der ChEMBL zusammengefasst. Diese kombinierten Bibliotheken wurden anschließend mithilfe des Clustering-Algorithmus StruClus (siehe Abschnitt II.1.3 und Schäfer und Mutzel (2017)) so unterteilt, dass in jedem Teil (Cluster) die Strukturen ähnlicher zueinander sind, als zu den Strukturen der anderen Cluster. StruClus wurde gewählt, da seine Laufzeit linear mit der Anzahl an Eingabemolekülen ansteigt (skaliert) und die kombinierte Datenbank für MCR 42 Millionen, für medchem 67 Millionen und für heterozyklen 24 Millionen Moleküle umfasst. Damit stellen die kombinierten Datenbanken sehr anspruchsvolle Datensätze dar, welche mit chemischen Standardclustering-Algorithmen nicht adäquat analysierbar sind.

Bevor CH*I*PMUNKs Sub-Bibliotheken mittels StruClus unterteilt wurden, erfolgte eine Evaluation des Clusteringalgorithmus von StruClus bezüglich seiner chemischen Aussagekraft und der Fähigkeit bioähnliche Moleküle (Beeinflussung derselben Zielstruktur) innerhalb von Clustern anzureichen. Dazu wurde der Datensatz von Stahl *et al.* (2005) (SBDS) verwendet (siehe Abschnitt II.3.1 und Stahl *et al.* (2005)) und diese Ergebnisse mit den Ergebnissen des häufig in der Chemieinformatik verwendeten Jarvis-Patrick-Clusterings basierend auf dem ECFP4 verglichen. Der Benchmarkdatensatz SBDS enthält 466 Liganden für neun Zielstrukturen. Für beide Clustering-Verfahren wurde untersucht, wie gut die generierte, Cluster-basierte Einteilung mit der natürlichen Einteilung in die neun Gruppen der Zielstrukturen (*ground truth*) übereinstimmt. Dazu wurde der F1-Score berechnet (siehe Abschnitt IV.2.8.1 und Larsen und Aone (1999)), der Werte zwischen null und eins ergibt. Ein Wert von eins bedeutet eine identische Einteilung, d. h. identische Cluster. Lägen in dieser Analyse identische Cluster vor, so würde dies eine Kongruenz von *ground truth*, entsprechend der Einteilung nach Zielstruktur, und der Einteilung durch StruClus bzw. Jarvis-Patrick bedeuten. StruClus erreichte einen Clustergrößen gewichteten/ungewichteten F1-Score von 0.41/0.22 und liegt damit im Bereich vom F1-Score des Jarvis-Patrick-Clusterings (F1-Score von 0.47/0.17). Dazu sind drei Punkte anzumerken. Erstens wurden die Parameter des Jarvis-Patrick-Clusterings optimiert, wohingegen die Parameter von StruClus weitestgehend den Standardwerten entsprachen (siehe auch Abschnitt IV.2.8.2). Die Ergebnisse stellen damit die maximal erreichbare Clusteringqualität bezogen auf den F1-Score des Jarvis-Patrick-Clusteringalgorithmus für diesem Datensatz dar. In der Praxis, in der die *ground truth* unbekannt ist, werden daher in der Regel schlechtere Ergebnisse erreicht. Die Abweichungen der Parametereinstellungen von den Standardeinstellung für StruClus entstanden durch die Anpassung an CH*I*PMUNK und sollten für den Vorversuch zu Gunsten einer besseren Vergleichbarkeit bereits angewendet werden. Zweitens konnte das Jarvis-Patrick-Clustering zwar einen leicht besseren gewichteten F1-Score erzielen, es erzeugte jedoch viele Singletons (Cluster mit nur einem Mitglied, d. h. Molekül). Drittens besitzt das Jarvis-Patrick-Clustering eine in Bezug auf die Anzahl der zu clusternden Verbindungen quadratische Laufzeit. Dementsprechend wäre eine Anwendung auf CH*I*PMUNK aufgrund der Datenbankgröße futil.

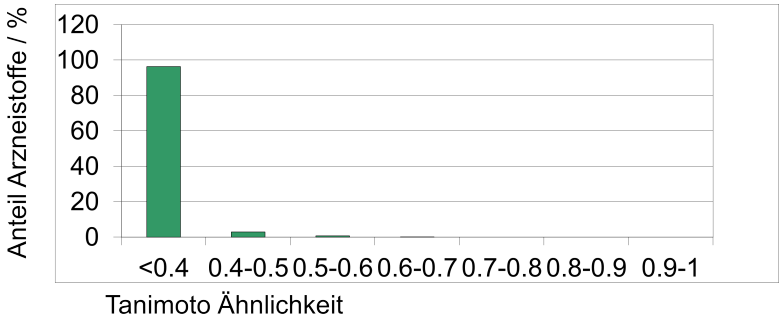
Zur Analyse der Clustering-Ergebnisse von StruClus auf den Sub-Bibliotheken von CH*I*PMUNK wurde der relative Anteil der Startbibliotheken (rot), der CH*I*PMUNK Sub-Bibliothek (grün) und von ChEMBL (blau) pro Cluster in einem Plot aufgetragen (Plots siehe Abbildung III.1.6, Gleichungen siehe Gleichungen IV.2.11 und IV.2.12). Falls eine Sub-Bibliothek einen neuartigen Subraum abdeckt, so müssten sich in dem Plot Cluster mit einem geringen Anteil an Molekülen aus den Startdatenbanken und der ChEMBL zeigen. Dies würde einem hohen Grünanteil entsprechen. Andererseits, falls die Datenbanken denselben chemischen Raum einnehmen, wäre eine zufällige Aufteilung in die Cluster (randomisiertes Clustering) zu erwarten. Innerhalb eines Clusters wären die Datenbanken entsprechend ihrer Größe vertreten und die Clusterzugehöhrigkeit würde nicht mit der Zugehörigkeit zu einer Datenbank korrelieren. Das Ergebnis wäre ein Plot mit ähnlichen Anteilen an allen drei Datenbanken in jedem Cluster. Zur Kontrolle wurde zu jeder Sub-Bibliothek ebenfalls ein randomisiertes Clustering erstellt. Alle randomisierten Plots sind sehr ähnlich, weshalb exemplarisch nur der Plot des randomisierten Clusterings der heterozyklen Sub-Bibliothek in Abbildung III.1.6d gezeigt ist. Im Vergleich zwischen den nicht randomisierten Clusterings (Abbildungen III.1.6a, III.1.6b, III.1.6c) und dem randomisierten Clustering (Abbildung III.1.6d) zeigt sich ein sehr deutlicher Unterschied. Daraus folgt, dass CH*I*PMUNK einen neuartigen chemischen Raum beinhaltet, in dem die MCR Sub-Bibliothek den größten Anteil (siehe Abbildung III.1.6c) einnimmt. Es ist anzunehmen, dass dieser neuartige Raum mit dem höheren Anteil an bRo5-Verbindungen, welcher ebenfalls für die MCR Sub-Bibliothek erzielt werden konnte, in Zusammenhang steht und sich in diesem voraussichtlich überproportional viele bRo5-Verbindungen befinden.

III.1.2.3. Vergleich von CH*I*PMUNK und DrugBank sowie iPPI

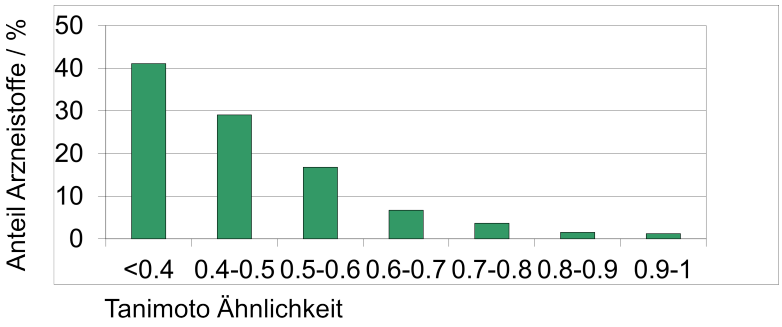
Anschließend an die Analyse des chemischen Raumes sollte eine genauere Betrachtung der Ähnlichkeit von Arzneistoffen aus der DrugBank und PPIIs aus der iPPI (Labbé *et al.* 2016; *iPPI-DB Inhibitors of Protein-Protein Interaction Database* 2018) zu CH*I*PMUNK-Verbindungen erfolgen. Dazu wurde der ECFP4 jeder Verbindung berechnet und die TC basierte Ähnlichkeit zwischen Arzneistoff bzw. PPII und CH*I*PMUNK-Verbindung je CH*I*PMUNK Sub-Bibliothek ermittelt. Die maximale Tanimoto-Ähnlichkeit (größter TC-Wert) jedes Arzneistoffs bzw. PPIIs zu einer CH*I*PMUNK-Verbindung je Sub-Bibliothek wurde gespeichert und ein Histogramm über alle Arzneistoffe (siehe Abbildung III.1.7) bzw. PPIIs (siehe Abbildung III.1.8) erstellt. Verbindungen mit hoher Ähnlichkeit zu DrugBank-Molekülen sind in der medchem und heterozyklen Sub-Bibliothek zu finden (siehe Abbildungen III.1.7a und III.1.7c). Zum Beispiel findet sich der Arzneistoff Citalopram (siehe Abbildung III.1.9a), ein Serotonin-Wiederaufnahmeinhibitor, in der heterozyklen Sub-Bibliothek wieder.



(a) heterozyklen

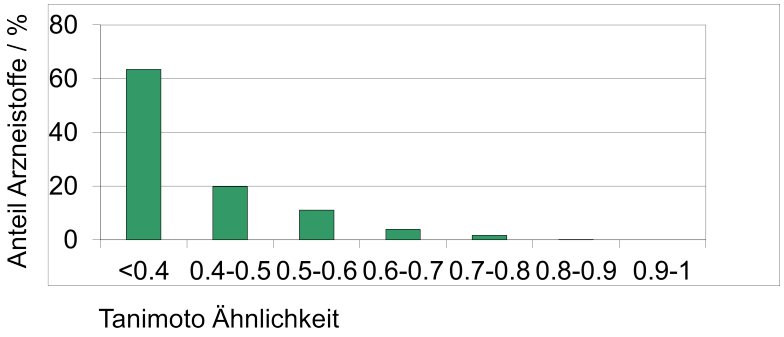


(b) MCR

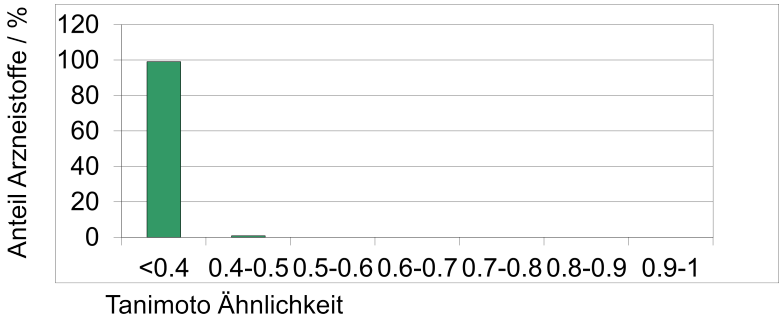


(c) medchem

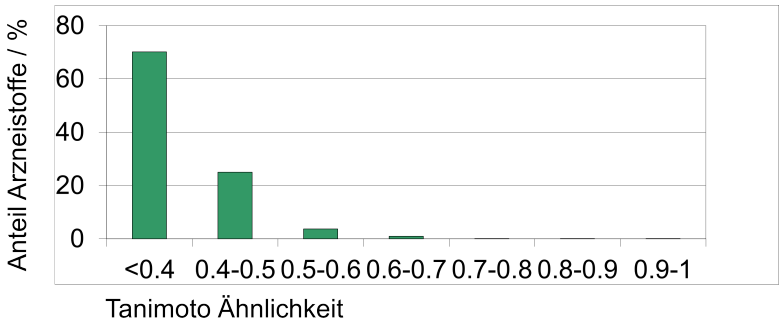
Abbildung III.1.7.: Analyse der Ähnlichkeit von CH*I*PMUNK Sub-Bibliotheken zur DrugBank: heterozyklen (a), MCR (b) und medchem (c). Auf der Abszisse sind die TC-Werte basierend auf dem ECFP4 zusammengefasst in definierte Bereiche aufgetragen gegen den prozentualen Anteil an Arzneistoffen auf der Ordinate.



(a) heterozyklen

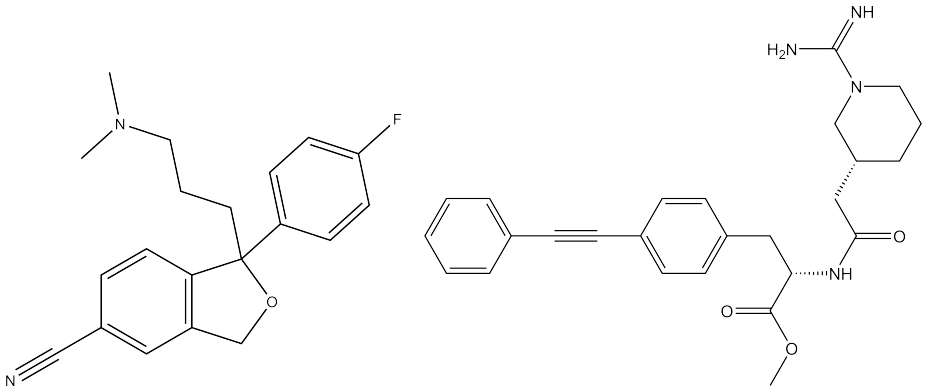


(b) MCR



(c) medchem

Abbildung III.1.8.: Analyse der Ähnlichkeit von CH*I*PMUNK Sub-Bibliotheken zur iPPI: heterozyklen (a), MCR (b) und medchem (c). Auf der Abszisse sind die TC-Werte basierend auf dem ECFP4 zusammengefasst in definierte Bereiche aufgetragen gegen den prozentualen Anteil an Arzneistoffen auf der Ordinate.



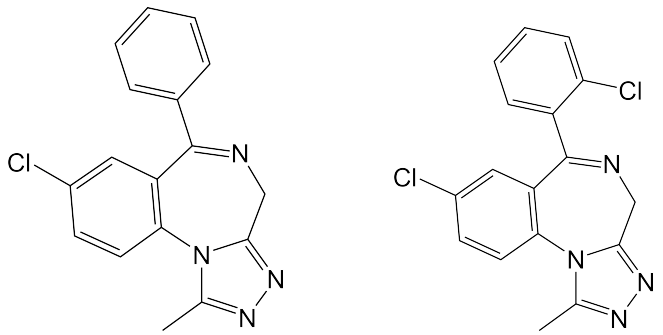
(a) Citalopram (b) iPPI 463

Abbildung III.1.9.: In CH*I*PMUNK enthaltener Arzneistoff (Citalopram (a)) und Protein-Protein-Interaktions-Inhibitor (iPPi 463 (b)).

Dementsprechend kann geschlussfolgert werden, dass der chemische Raum von CH*I*PMUNK Bereiche des chemischen Raumes von Arzneistoffen einschließt. Zusammen mit der Tatsache, dass CH*I*PMUNK dennoch ebenfalls einen hohen Anteil an neuem chemischen Raum abdeckt, kann erwartet werden, dass CH*I*PMUNK noch nicht entdeckte, zukünftige Arzneistoffe beinhaltet. Darüberhinaus konnte eine angemessene Ähnlichkeit von TC $>$ 0,8 und eine hohe Ähnlichkeit von TC $>$ 0,9 zu bekannten PPIIs für Verbindungen der CH*I*PMUNK heterozyklen bzw. medchem Sub-Bibliothek festgestellt werden (siehe Abbildungen III.1.8a und III.1.8c). In der medchem Sub-Bibliothek konnte zum Beispiel der Interleukin-2-Binder mit der iPPI ID 463 (siehe Abbildung III.1.9b) wiedergefunden werden. Es ist anzumerken, dass im Vergleich zur ChEMBL, welche zahlreiche Zielstrukturen umfasst, die iPPI nur auf 18 verschiedenen Zielstrukturen beruht. Ein aussagekräftiger Vergleich der chemischen Räume von iPPI und CH*I*PMUNK ist somit erschwert. Infolge der aktuellen Entwicklung sollte die Anzahl an PPIIs rasch anwachsen und zukünftig einen tieferen Einblick in den chemischen Raum von PPIIs gewähren.

III.1.2.4. Anwendungsbeispiel

CH*I*PMUNK kann dazu verwendet werden, um in einem virtuellen Screening ähnliche Moleküle zu einem Molekül von Interesse (Query) zu identifizieren. Der gesamte Cluster, in dem sich das Query-Molekül befindet, kann als zielgerichtete Bibliothek genutzt werden (*targeted library*). Dadurch wird der Suchraum eingeengt und somit die Anwendung von rechenaufwändigen Methoden ermöglicht. Von bedeutsamen Nutzen ist, dass der Cluster neben dem Query sowohl ähnliche neue Moleküle als auch kaufbare Moleküle beinhalten kann und darüberhinaus eine erste Einschätzung des Bioaktivitätsspektrums durch ähnliche Moleküle mit vorhandenen Bioaktivitätsdaten erfolgen kann.



(a) Alprazolam (b) Triazolam

Abbildung III.1.10.: Die Suche in CH*I*PMUNK nach dem Arzneistoff Alprazolam (a) liefert Triazolam (b).

Als Beispiel wurde mithilfe eines KNIME-Workflows (siehe Abschnitt IV.2.9) in der heterozyklen Sub-Bibliothek nach Alprazolam (siehe Abbildung III.1.10a), einem Anxiolytikum, gesucht. Ähnliche Moleküle werden in dem Cluster mit der ID 968, welcher das ChEMBL Molekül Triazolam (ChEMBL646, siehe Abbildung III.1.10b) enthält, gefunden. Triazolam ist ebenfalls ein anxiolytischer $GABA_A$-Rezeptor-Modulator. Die Einteilung in Cluster durch StruClus ist ein besonderes und herausragendes Merkmal von CH*I*PMUNK. Für Optimierungsprojekte kann CH*I*PMUNK dank der Cluster ebenfalls dienen und auf diese Weise innovative Alternativen mit verbessertem pharmakokinetischen und/oder pharmakodynamischen Profil liefern.

III.1.3. Zusammenfassung

CH*I*PMUNK ist eine besonders wertvolle *in silico* Datenbank, da sie mit ihren drei Sub-Bibliotheken mit insgesamt mehr als 95 Millionen Molekülen sowohl den Deskriptor-Raum von Protein-Ligand- und Protein-Protein-Interaktions-Inhibitoren als auch von Liganden für bisher als schwer durch (niedermolekulare) Verbindungen zugänglich geltende (in der Literatur als „undruggable“, „difficult to drug“ oder „yet to be drugged“ bezeichnet (Dang *et al.* 2017)) Proteine (bRo5) umfasst. Außerdem wurde unter Berücksichtigung der Synthetisierbarkeit ein neuartiger chemischer Raum generiert, welcher auf innovative Arzneistoffe und *chemical probes* hoffen lässt. Dies sollte durch die Auswahl der Reaktionen erreicht werden und konnte zum einen durch einen akzeptablen Anteil an SA-Scores größer null (mindestens 20 % mit Ausnahme der MCR Sub-Bibliothek aus CoMol-Edukten) und zum anderen durch die Anwesenheit bereits bekannter Arzneistoffe wie Citralopram oder PPIIs wie iPPI 463 in CH*I*PMUNK gezeigt werden. *Chemical Probes* sind niedermolekulare Verbindungen, die aufgrund ihrer hohen Potenz und Selektivität zur Untersuchung ihres Zielproteins, z. B. dessen physiologischer Funktion, und zur Identifizierung sowie Entwicklung weiterer Modulatoren des Proteins, genutzt werden können (Arrowsmith *et al.* 2015).

Der bereits bekannte chemische Raum von CH*I*PMUNK wurde aufgeteilt und der entsprechende bioaktive Raum den Teilen (Clustern) zugeordnet, sodass eine erste Abschätzung der potentiellen Zielstruktur ermöglicht wird. Zudem wird die Erstellung einer Target-spezifischen Bibliothek erleichtert. Zusammenfassend ist CH*I*PMUNK eine innovative und wichtige *in silico* Bibliothek für virtuelle Screeningansätze, zur Optimierung und Diversifizierung eines Hits sowie zur Ideengenerierung für neuartige Arzneistoffe. Außerdem liefert sie einen Ansatzpunkt für Targetspekulationen, z. B. für Hits aus phänotypischen Screens. Eine weitere Anwendung besteht in der Nutzung von CH*I*PMUNK als Benchmarkdatensatz für Algorithmen für große Datenmengen (Big Data). Big Data Algorithmen werden dringend benötigt, um die stetig wachsenden Daten aufzuarbeiten und die enthaltenen Informationen zu extrahieren. Ihre Evaluierung und Validierung kann durch eine freiverfügbare Bibliothek wie CH*I*PMUNK erleichtert werden. CH*I*PMUNK steht unter dem Link *ccb.tu-dortmund.de/ag-koch/chipmunk* zur Verfügung. In den folgenden Kapiteln werden einige neue Methoden vorgestellt, die final auf CH*I*PMUNK getestet und angewendet werden könnten. Es ist davon auszugehen, dass CH*I*PMUNK, als innovative Moleküldatenbank, einerseits die Identifizierung neuer Wirkstoffe fördert und erleichtert sowie andererseits in Zukunft zur Entwicklung neuer, effizienter Algorithmen beiträgt.

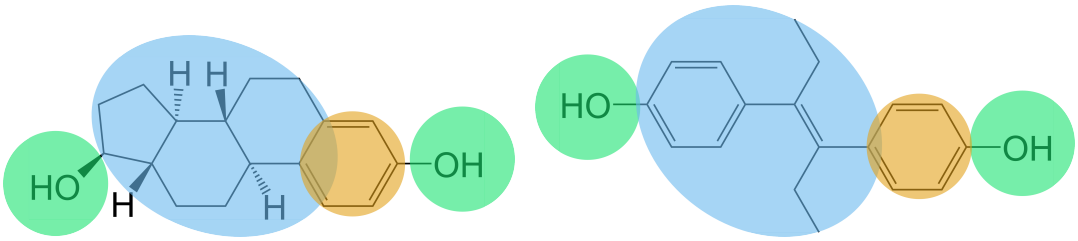
Quintessenz:

- die *in silico* Datenbank CH*I*PMUNK mit mehr als 95 Millionen Verbindungen wurde erstellt
- CH*I*PMUNK deckt den chemischen und physikochemischen Raum von Protein-Ligand- und Protein-Protein-Interaktions-Inhibitoren ab
- das StruClus-Clustering konnte erfolgreich validiert und angewendet werden
- definierte Cluster ermöglichen die Abschätzung des Targetraums innovativer CH*I*PMUNK-Verbindungen

III.2. Die Atomtypen- und Pharmakophor-Ebene der Ähnlichkeit

III.2.1. Einleitung

Im Gegensatz zur Kennzeichnung der Atome eines Moleküls bzw. der Knoten eines Molekülgraphen mit den Elementbezeichnungen bietet die Kennzeichnung mit dem Atomtyp die Möglichkeit auch den Einfluss der Nachbarschaft auf die Eigenschaften des jeweiligen Atoms zu erfassen. Es gibt verschiedene Atomtypen-Definitionen, z. B. von Sybyl, *general AMBER force field* (GAFF) (Wang *et al.* 2004; Wang *et al.* 2006) oder *file conversion* (fconv) (Neudert und Klebe 2011). Als Ausgangspunkt der Atomtypen-Ebene wurde in dieser Arbeit die fconv-Atomtypendefinition verwendet, da diese eine feinere Untergliederung der Atomumgebungen liefert. Jedoch ist zu erwähnen, dass die GAFF-Atomtypen ebenfalls wie die fconv-Atomtypen angepasst werden können, weshalb GAFF für zukünftige Untersuchungen mit neu definierten Atomtypen durchaus in Betracht gezogen werden sollte. Die Umgebung eines Atoms kann dessen Eigenschaften wesentlich beeinflussen. Um die Ähnlichkeit von Atomen basierend auf ihrer Nachbarschaft und damit verbundenen physikochemischen Eigenschaften zu betrachten, werden mehrere ähnliche Atomtypen in einer Atomtypengruppe zusammengefasst. Die Atome werden anschließend mit der Atomtypengruppe gekennzeichnet. Die Ergebnisse der Ebene der Ähnlichkeit basierend auf Atomtypengruppen werden in diesem Kapitel ausgewertet und diskutiert. Zum Vergleich werden zusätzlich die Ergebnisse von Ähnlichkeitssuchen basierend auf Atomytypen- und Element-Kennzeichnungen analysiert.
Eine offizielle *International Union of Pure and Applied Chemistry* (IUPAC) Definition eines Pharmakophors befindet sich in Wermuth *et al.* (1998). Ein Pharmakophor besteht aus der dreidimensionalen Anordnung von sterischen und elektronischen Eigenschaften (pharmakophore Eigenschaften), die eine Wechselwirkung mit einer spezifischen biologischen Zielstruktur erlaubt. Diese Wechselwirkung resultiert in einer modulierten biologischen Antwort der biologischen Zielstruktur. Bei pharmakophoren Eigenschaften handelt es sich nicht um funktionelle Gruppen, sondern um vom Molekül abstrahierte Eigenschaften wie Wasserstoffbrückenakzeptor, Wasserstoffbrückendonor, geladene Gruppe, hydrophobe oder aromatische Region. Ein Beispiel ist in Abbildung III.2.1 gezeigt.



(a) Estradiol (b) Diethylstilbestrol

Abbildung III.2.1.: Beispiel für zwei Moleküle mit gleichem Pharmakophor: Estradiol (a) und Diethylstilbestrol (b). Beide sind Agonisten des Estrogenrezeptors α. Die Farben codieren die pharmakophoren Eigenschaften, grün: Doneptor, blau: hydrophob und orange: aromatisch.

In einer pharmakophoren Eigenschaft können mehrere Atome zusammengefasst sein, z. B. ein aromatischer Ring kann als ein Punkt bzw. Knoten dargestellt werden, welcher räumlich z. B. im Mittelpunkt des ursprüglichen Rings lokalisiert sein kann. Damit wird der Molekülgraph vereinfacht und man spricht von einem reduzierten Molekülgraphen (siehe Abschnitt II.1.2 und Beispiele in Abbildungen II.0.2d, II.0.2e und II.0.2f). Das Pharmakophorkonzept ist ein wichtiges Konzept in der computergestützen Wirkstofforschung. Da es Ähnlichkeiten zwischen distinkten Atomen oder Atomtypen über deren sterische und elektronische Eigenschaften aufzeigt und die Interaktion mit der Zielstruktur als einen essentiellen Bestandteil beinhaltet, soll es in Form der Pharmakophor-Ebene der Ähnlichkeit ebenfalls in diesem Kapitel untersucht werden.
Zur Untersuchung des Einflusses der Verwendung von Knotenkennzeichnungen (Knotenlabeln) basierend auf Element, Atomtyp, Atomtypengruppe und pharmakophoren Eigenschaft auf die Performanz von Ähnlichkeitssuchen mittels BBPMCS (für Details siehe II.1.3 und Droschinsky, Kriege und Mutzel (2017)) wurde der Datensatz von O'Boyle und Sayle (OBDS) (siehe Abschnitt II.3.2 und O'Boyle und Sayle (2016)) verwendet. Für entsprechende Untersuchungen der Performanz des StruClus-Clustering-Algorithmus (für Details siehe Abschnitt II.1.3 und Schäfer und Mutzel (2017)) wurde der Datensatz von Stahl *et al.* (SBDS) (siehe Abschnitt II.3.1 und Stahl *et al.* (2005)) genutzt.

III.2.2. Ergebnisse und Diskussion

III.2.2.1. fconv-Atomtypen und fconv-Atomtypengruppen

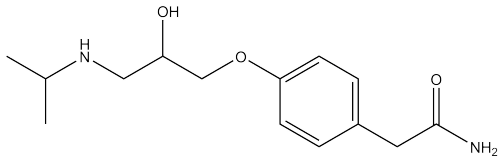
Die fconv-Atomtypen wurden in acht Gruppen eingeteilt: (i) Gruppe 0: nicht aromatischer Ring, (ii) Gruppe 1: Wasserstoffbrückenakzeptor, (iii) Gruppe 2: Wasserstoffbrückendonor, (iv) Gruppe 3: aromatisch, (v) Gruppe 4: hydrophob, (vi) Gruppe 5: Halogen, (vii) Gruppe 6: negativ geladen und (viii) Gruppe 7: positiv geladen. Eine Zuordnung der fconv-Atomtypen zu den acht Gruppen befindet sich im Anhang unter A.1.

Proof of Concept Studie

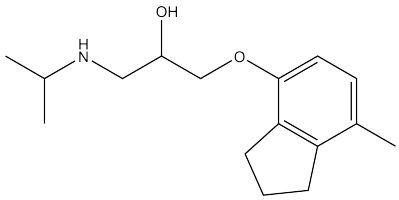
Zu Beginn der Arbeit an diesem Teilprojekt wurde eine Proof of Concept Studie durchgeführt. Ziel war eine Gegenüberstellung der State of the Art Methode einer Ähnlichkeitssuche (Tanimoto-Koeffizient basierte Ähnlichkeit beruhend auf dem ECFP4 (Rogers und Hahn 2010), ECFP4-TC) (Duvenaud *et al.* 2015) mit der in dieser Arbeit neu entwickelten Ähnlichkeitssuche mittels BBPMCS und fconv-Atomtypengruppen (MSS-Ansatz). Dazu wurde in der ChEMBL (Bento *et al.* 2014) nach dem Arzneistoff Atenolol (siehe Abbildung III.2.2a) gesucht und die gefundenen Moleküle mit einem MSS-Score bis 0,65 unter Verwendung der Ähnlichkeitsfunktion 1 (siehe Tabelle III.2.4) (MSS-Hits, ECFP4-TC-Hits analog) wurden analysiert. Atenolol wurde als Referenz gewählt, da für Atenolol einige ähnliche Moleküle in der ChEMBL zu erwarten waren. Die Verteilung der Scores befindet sich im Anhang unter A.1. Atenolol ist ein β_1-Adrenozeptor-Antagonist (UniProt-ID: P08588 (The UniProt Consortium 2017; *UniProt* 2018)) zur Behandlung von z. B. Hypertonie und weiteren kardiovaskulären Erkrankungen. Eine Übersicht der Hits befindet sich in Tabelle III.2.1. Der β_1-Adrenozeptor der Ratte besitzt die UniProt-ID P18090. Als erstes kann festgestellt werden, dass die gefundenen Hits einen sehr ähnlichen Targetraum abdecken. Folglich ist der MSS-Ansatz in der Lage, bioähnliche Moleküle als ähnlich zu bewerten, indem sie einen hohen MSS-Score erhalten. Damit wird eine erste wichtige Voraussetzung zur Anwendung von Graph-basierten Ähnlichkeitsmethoden in Kombination mit der Ebene der Atomtypengruppen-Ähnlichkeit erfüllt. Nichtsdestotrotz können die nach MSS-Ansatz ähnlichen Moleküle auf demselben Target unterschiedliche Wirkungen besitzen, so ist Atenolol ein β_1-Adrenozeptor-Antagonist, wohingegen Prenalterol ein β_1-Adrenozeptor-Agonist ist.

Des Weiteren ist anhand des unterschiedlichen Rankings deutlich erkennbar, dass sich die beiden Methoden (ECFP4-TC und MSS-Ansatz) hinsichtlich ihrer Bewertung der Ähnlichkeit unterscheiden und somit unterschiedliche Moleküle als Hits hervorbringen. Dies ist in guter Übereinstimmung mit den Ergebnissen von Raymond und Willett (2002).

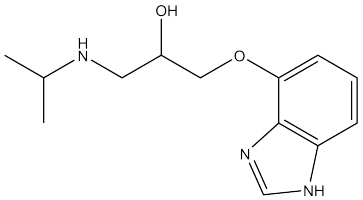
Besonders markant ist dies für die MSS-Hits CHEMBL26917 (Rang 14, Struktur siehe Abbildung III.2.2b) und CHEMBL3228931 (Rang 16, Struktur siehe Abbildung III.2.2c). Beide besitzen einen ECFP4-TC von kleiner bzw. gleich 0,4, somit würden diese Moleküle in der Regel nicht zur weiteren Untersuchung ausgewählt werden.



(a) Atenolol



(b) CHEMBL26917



(c) CHEMBL3228931

Abbildung III.2.2.: Strukturen der Hits der MSS-Suche nach Atenolol (a), welche nicht durch ECFP4-TC gefunden werden konnten (bis zu einem TC von größer 0,4): CHEMBL26917 (Rang 14) (b) und CHEMBL3228931 (Rang 16) (c).

Unterstützt wird dieses Ergebnis durch eine Analyse der Scores aller Verbindungen des ChEMBL-Datensatzes. Für jedes Molekül dieses Datensatzes wurde der ECFP4-TC-Score basierend auf den Sybyl-Atomtypen gegen den MSS-Score aufgetragen (siehe Anhang A.2). Der Korrelationskoeffizient (r) beträgt gerundet nur 0,32. Für die Auftragung der ECFP4-TC-Scores basierend auf fconv-Atomtypengruppen gegen die MSS-Scores beträgt der Korrelationskoeffizient gerundet 0,4. Nichtsdestotrotz handelt es sich um eine statistisch signifikante Korrelation, da der p-Wert in beiden Fällen Null ist und somit die Nullhypothese, dass die Steigung Null ist, abgelehnt werden muss.

Aus dem Vergleich vom ECFP4-TC Rang und ECFP4-TC Rang fortlaufend in Tabelle III.2.1 kann abgeleitet werden, dass beim ECFP4-TC viele Platzierungen mehrfach vergeben sind, da sich der TC nicht unterscheidet. Der ECFP4-TC liefert unter den 1000 am höchsten bewerteten Molekülen 239 verschiedene Scores. Der MSS-Ansatz erzielt hingegen unter den 1000 am höchsten bewerteten Molekülen 431 verschiedene Scores. Dies bedeutet, dass beim ECFP4-TC nur circa jedes vierte Molekül einen eigenen Score erhält. Beim MSS-Ansatz erhält hingegen circa jedes zweite Molekül einen distinkten Score. Damit erzielt der MSS-Ansatz bemerkenswerterweise circa 1,8 mal so viele verschiedene Scores wie der ECFP4-TC. Im Hashing-Prozess bei der Generierung des ECFP4 kann es zu Kollisionen kommen. Dies bedeutet, dass zwei ursprünglich verschiedene Werte auf denselben Wert abgebildet werden. Um die Wahrscheinlichkeit einer solchen Kollision zu verringern, kann der Raum, in den abgebildet wird, vergrößert werden. Dazu wird die Anzahl der zur Verfügung stehenden Bits (Fingerprintlänge) erhöht. Es konnte für die Proof of Concept Studie jedoch gezeigt werden, dass Kollisionen nur einen sehr geringen Einfluss haben und den Unterschied nicht erklären können.

Mit einer Fingerprintlänge von 16384 bit konnten unter den 1000 am höchsten bewerteten Molekülen nur 255 einzigartige Scores erreicht werden. Folglich konnten nur 16 weitere einzigartige Scores durch eine Erhöhung der Fingerprintlänge von 1048 bit auf 16384 bit beobachtet werden und in diesen Fällen kann eine Kollision als Grund für denselben Score angenommen werden. Auf eindrucksvolle Weise wird diese Beobachtung durch den Plot der fortlaufenden Scores auf der Abszisse gegen die einzigartigen Scores auf der Ordinate in Abbildung III.2.3 bestätigt. Es ist ein deutlicher Unterschied zwischen den beiden ECFP4-TC-Methoden (grüne und blaue Punkte) und dem MSS-Ansatz (magentafarbene Punkte) zu erkennen. Zum einen ist die, durch die magentafarbenen Punkte entstehende, Linie deutlich länger und zum anderen ist die Linie sowohl zu Beginn als auch am Ende deutlich steiler. Dies bedeutet, dass einerseits die Moleküle besser auf dieser Linie verteilt liegen und andererseits, besonders bei hohen und bei niedrigen Scores, die diskriminative Stärke des MSS-Ansatzes liegt. Sowohl der hohe als auch der niedrige Scorebereich sind von praktischer Bedeutung für virtuelle Screeningansätze bzw. zur Erstellung von diversen Substanzbibliotheken. Der Unterschied der Methoden im maximal erreichten Rang liegt darin begründet, dass durch die verschiedene Repräsentation der Moleküle unterschiedlich viele Moleküle als Duplikate erkannt und anschließend aussortiert werden.

Daraus kann geschlussfolgert werden, dass der MSS-Ansatz besser zwischen Molekülen unterscheiden kann, obwohl für den MSS-Score weniger Nachkommastellen betrachtet wurden als für die ECFP4-TC-Methoden. Damit konnte die Annahme, dass die Verwendung der Graphstruktur eine genauere Repräsentation von Molekülen als z. B. Fingerprints erlaubt, gestützt werden. Mit dem in der Abbildung III.2.4 gezeigten Beispiel soll verdeutlicht werden, dass die erhöhte diskriminative Kraft nützlich und biologisch relevant ist. CHEMBL6920 (Struktur siehe III.2.4b) wird durch den ECFP4-TC als gleich ähnlich zu Atenolol (III.2.4a) bewertet wie CHEMBL3272440 (III.2.4c) (beide erhalten einen ECFP4-TC von rund 0,446809). Der MSS-Ansatz bewertet CHEMBL6920 hingegen mit einer deutlich höheren Ähnlichkeit (0,702128) zu Atenolol als CHEMBL3272440 (0,42953). Allerdings gibt es beim MSS-Ansatz wiederum Moleküle, die dieselben Scores (0,702128 bzw. 0,42953) erzielen (Strukturen und Scores siehe Anhang A.4). Beide Moleküle sind Antagonisten für den β_2-Adrenozeptor des Hausrinds bzw. Hausmeerschweins (Uniprot-IDs: Q28044 bzw. Q8K4Z4) (siehe ChEMBL-Einträge (Bento *et al.* 2014)). CHEMBL6920 besitzt, verglichen mit Rezeptoren der Luftröhre, eine höhere Aktivität (niedrigerer Kd) an Rezeptoren der rechten Arterie. Für CHEMBL3272440 gilt das Umgekehrte. Die β_1-Adrenozeptoren sind der vorherrschende β-Adrenozeptortyp des Herzens, β_2-Adrenozeptoren sind vornehmlich in der glatten Muskulatur lokalisiert. Dies könnte ein Hinweis darauf sein, dass CHEMBL6920, wie Atenolol, eine β_1-Präferenz besitzt, wohingegen CHEMBL3272440 vermutlich eine β_2-Präferenz aufweist.

Faszinierend ist in diesem Zusammenhang, dass es eine Balance zwischen chemisch-molekularer Genauigkeit bzw. Diversität und biologischer *Fuzziness* (Unschärfe) bzw. Ähnlichkeit gibt. Dies meint, dass chemisch verschiedene Moleküle nichtsdestotrotz ähnliche biologische Wirkungen erzielen können.

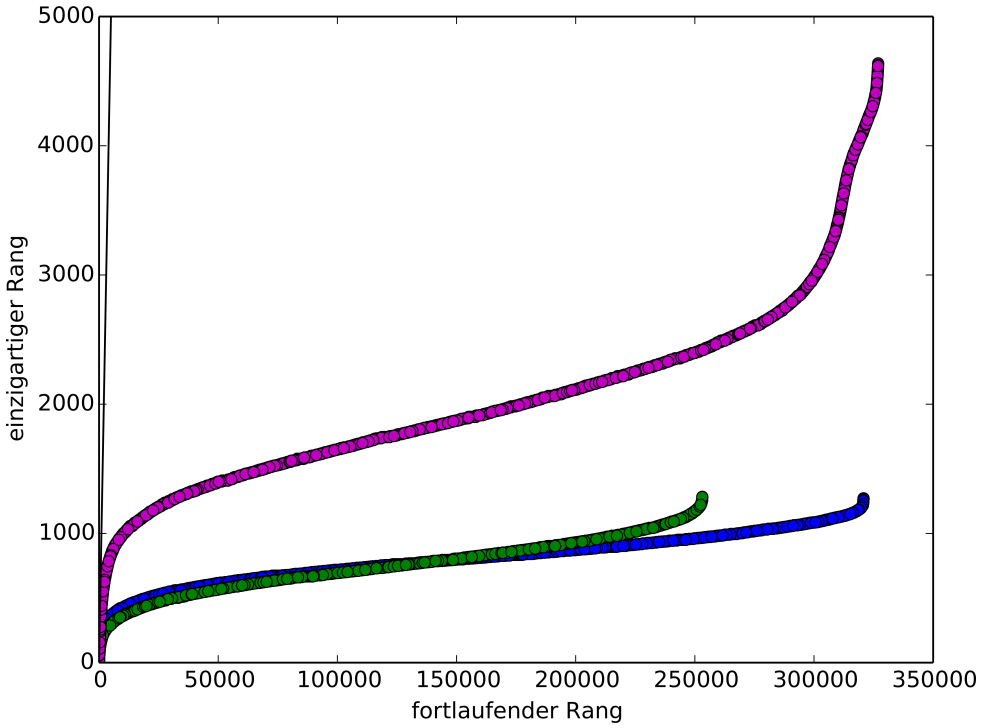
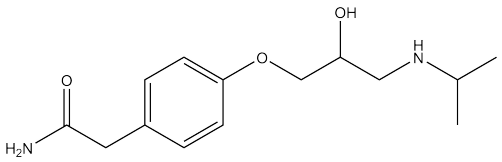
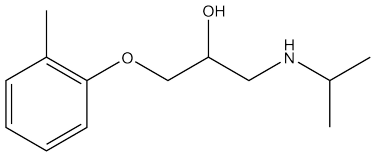


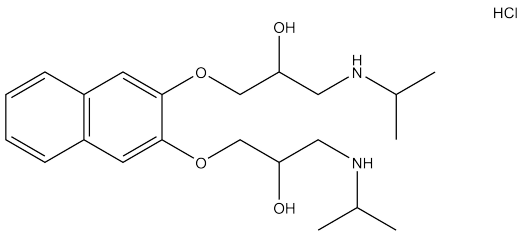
Abbildung III.2.3.: Vergleich der Ränge in der Proof of Concept Studie für verschiedene Methoden. Die einzigartigen Ränge (Ordinate) sind gegen die fortlaufenden Ränge (Abszisse) der MSS-Suche (magenta), der ECFP4-TC-Suche mit Sybyl-Atomtypen (blau) und der ECFP4-TC-Suche mit fconv-Atomtypengruppen (grün) aufgetragen. Hätte jedes Molekül einen distinkten Score so würde sich die in schwarz gezeichnete Gerade ergeben.



(a) Atenolol



(b) CHEMBL6920
ECFP4-TC: 0,446808510638298
MSS: 0,702128



(c) CHEMBL3272440
ECFP4-TC: 0,446808510638298
MSS: 0,42953

Abbildung III.2.4.: Strukturen der Hits der MSS-Suche nach Atenolol (a), welche nicht durch den ECFP4-TC unterschieden werden können: CHEMBL6920 (b) und CHEMBL3272440 (c).

So wird das Molekül zwar durch die Verwendung der Graphstruktur genauer respräsentiert, jedoch wird durch die Verwendung der Atomtypengruppen eine definierte Unschärfe eingeführt, um somit eine bessere Abbildung der biologische Ähnlichkeit zu erlangen. Dies ist gut erkennbar bei der Betrachtung der Verteilung der Scores (siehe Anhang A.1). Für den ECFP4-TC basierend auf fconv-Atomtypengruppen (grüne Balken) verschiebt sich die Score-Verteilung hin zu höheren Werten, d. h. größerer Ähnlichkeit. Die Score-Verteilung des MSS-Ansatzes ist ebenfalls in Bezug auf den ECFP4-TC mit Sybyl-Atomtypen leicht zu höheren Werten verschoben. Jedoch konnte in den zuvor präsentierten Ergebnissen gezeigt werden, dass der MSS-Ansatz eine größere diskriminative Kraft besitzt als die ECFP4-TC-Methoden. Zudem erzielt ein ECFP4-TC kombiniert mit fconv-Atomtypengruppen nur 176 verschiedene Scores unter den ersten 1000 am höchsten bewerteten Molekülen. Es erhält folglich nur circa jedes sechste Molekül einen einzigartigen Score. Die Ergebnisse für den ECFP4-TC kombiniert mit fconv-Atomtypengruppen befinden sich im Anhang A.6. Sie dienen lediglich der zusätzlichen Bestätigung, da es sich nicht um den State of the Art, sondern eine Abwandlung dieses Standards handelt. Folglich wurde primär auf den Vergleich des neuen MSS-Ansatzes mit dem State of the Art ECFP4-TC mit Sybyl-Atomtypen fokussiert.

Element-Ebene

Im Anschluss an den erfolgreichen Abschluss der Proof of Concept Studie wurde eine extensive Testung auf dem OBDS Benchmark (siehe O'Boyle und Sayle (2016) und Abschnitt II.3.2) vorgenommen. Zuerst sollte ein Schritt zurückgegangen werden und die Performanz des BBPMCS ohne Atomtypen getestet werden. Dazu wurden die Atome im Molekülgraph nur mit ihrem Element gekennzeichnet.

Rang MSS	Trivialname und ChEMBL-ID	Target UniProt-ID (Organismus)	Rang ECFP4-TC	Rang ECFP4-TC fortlaufend
1	Atenolol	P08588	1	1
2	CHEMBL143467	P18090 (Ratte/Rattus norvegicus)	2	2
3	CHEMBL7154	P18090 (Ratte/Rattus norvegicus)	11	12
4	Prenalterol (CHEMBL1160714)	P08588	7	8
5	CHEMBL3247307	P18090 (Ratte/Rattus norvegicus)	4	4
6	Metoprolol (CHEMBL13 bzw. CHEMBL1256957)	P08588	6	6 und 7
7	CHEMBL1159888	Q28044* (Kuh/Bos taurus)	45	65
8	Esmolol (CHEMBL768)	P08588	3	3
9	Toliprolol (CHEMBL67096)	P07550* B0FL73 (Meerschwein/Cavia porcellus)	17	20
9	CHEMBL6920	Q28044* (Kuh/Bos taurus)	32	45
10	CHEMBL3228930	B0FL73 (Meerschwein/Cavia porcellus)	30	41
11	CHEMBL1159891	Q28044* (Kuh/Bos taurus)	37	51
12	CHEMBL3247301	P18090 (Ratte/Rattus norvegicus)	5	5
12	Practolol (CHEMBL6995)	P08588	8	9
13	CHEMBL6939	P18090 (Ratte/Rattus norvegicus) B0FL73 (Meerschwein/Cavia porcellus)	13	14
14	CHEMBL26917	P08588	54	95
14	CHEMBL6902	P18090 (Ratte/Rattus norvegicus)	14	15
15	CHEMBL1159890	Q28044* (Kuh/Bos taurus)	45	66
16	CHEMBL3228931	B0FL73 (Meerschwein/Cavia porcellus)	51	82
17	CHEMBL31608	B0FL73 (Meerschwein/Cavia porcellus)	9	10

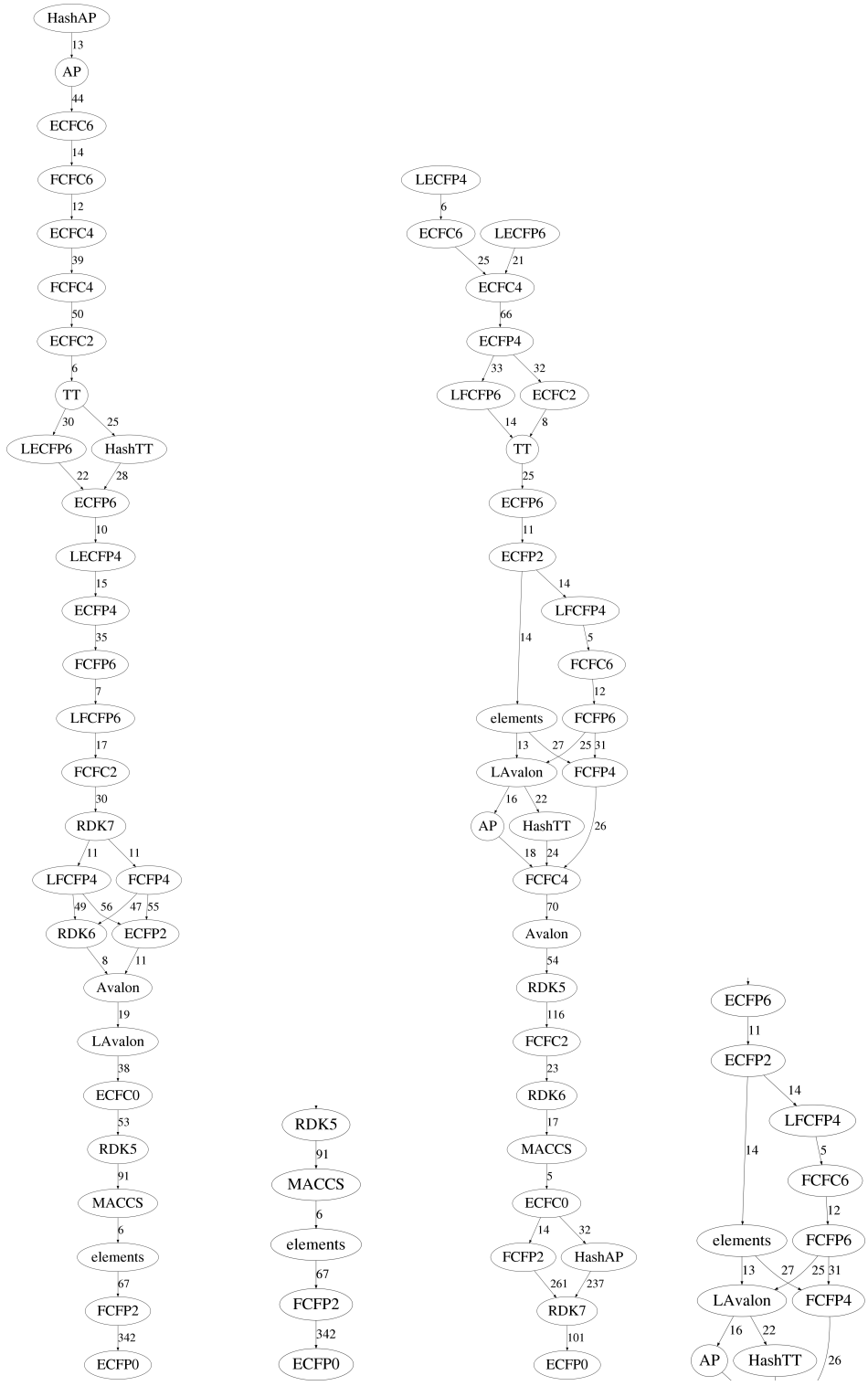
Tabelle III.2.1.: Ergebnisse der Proof of Concept Studie zum Vergleich von ECFP4-TC (State of the Art) und MSS-Ansatz. Für nicht humane Targets, wurde der Organismus in Klammern angegeben. * Target ist ein β_2-Adrenozeptor.

Das Ergebnis befindet sich in Abbildung III.2.5. Der BBPMCS in Kombination mit Elementlabeln hat eine bessere Performanz auf dem Multi-Assay-Datensatz im Vergleich zum Single-Assay-Datensatz. Da die Präsentation aller Ergebnisse schnell unübersichtlich wird und sich die Ergebnisse der Fingerprints nicht verändern, werden in den folgenden Abbildungen nur die Bereiche gezeigt, in denen sich die neuen Ansätze befinden. Alle dazugehörigen vollständigen Abbildungen können dem Anhang unter dem Abschnitt A.8 entnommen werden.

fconv-Atomtypen und fconv-Atomtypengruppen

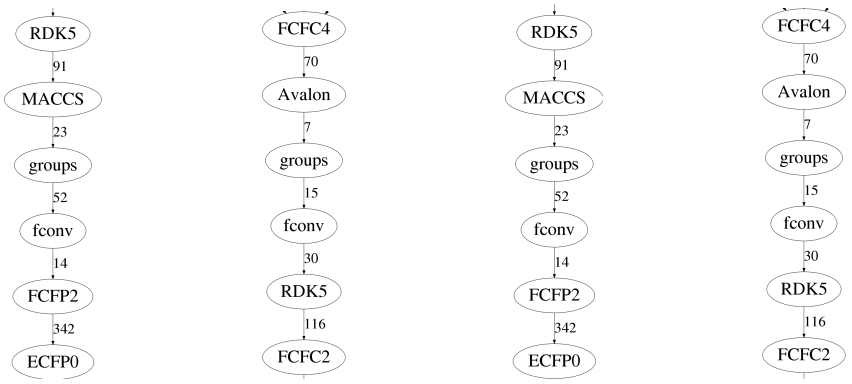
Der OBDS enthält einige große Moleküle (mehr als 1000 Atome), welche zusätzlich nicht außenplanare Graphen sind. Demzufolge muss zur Berechnung der Ähnlichkeit für diese Moleküle ein Algorithmus mit exponentieller Worst Case Laufzeit benutzt werden. Dies führte zu einem erheblichen Anstieg der Laufzeit. Zur Reduktion der Laufzeit auf eine praktikable Zeit, wurde ein Cutoff-Parameter (ct) eingeführt, welcher die Zeit in Millisekunden (ms) pro Clique angibt nach der die Suche nach der größten Clique abgebrochen wird. Die bis dahin beste gefundene Lösung wird anschließend verwendet. In einem Vortest sollten für den Parameter ct Werte von 1000 ms und 2000 ms getestet werden. Wie in Abbildung III.2.6 erkennbar ist, gibt es keinen qualitativen Unterschied zwischen den beiden Cutoffs, trotzdem wurde in den folgenden Experimenten ein Cuttoff von 2000 Millisekunden verwendet. Im Vergleich zu der Kennzeichnung mit Elementlabel (Abbildung III.2.5) schneiden die Kennzeichnungen mit fconv-Atomtypen sowie mit fconv-Atomtypengruppen in beiden Assays schlechter ab. Allerdings schneiden in beiden Assays die fconv-Atomtypengruppen besser ab als die reinen fconv-Atomtypen. Eine mögliche Interpretation ist, dass die fconv-Atomtypen eine zu genaue molekulare Repräsentation liefern, welche sich nicht in gleicher Auflösung in der biologischen Aktivität widerspiegelt. Durch die Verwendung der fconv-Atomtypengruppen wird diese Genauigkeit aufgelockert und schafft Raum für biologische *Fuzziness*. Die folgenden zwei Versuche fokussieren sich aus diesem Grund auf die fconv-Atomtypengruppen.

In zwei anschließenden Versuchen wurden dem BBPMCS-Algorithmus (siehe Abschnitt II.1.3) zusätzlich zwei verschiedene Ähnlichkeitsmatrizen übergeben. Durch die Ähnlichkeitsmatrizen ist es möglich, nicht nur zwischen identischer Gruppe (kann gemapped werden) und nicht identischer Gruppe (kann nie gemapped werden) zu unterscheiden, sondern die Ähnlichkeit der Grupppen weitergehend zu differenzieren. Zum Beispiel ist bei einem Mapping der hydrophoben Gruppe (4) auf die aromatische Gruppe (3) eher mit dem Erhalt der biologischen Aktivität zu rechnen als bei einem Mapping der hydrophoben Gruppe (4) auf die negativ geladene Gruppe (6). Die Ähnlichkeitsmatrix 1 ist in Tabelle III.2.2 gezeigt. Dabei entspricht die Gruppe aus der ersten (linken) Spalte der gesuchten Gruppe, dementsprechend einer Gruppe, die im Eingabemolekül vorhanden ist. Die dazugehörigen Ähnlichkeitswerte befinden sich in den jeweiligen Reihen, dabei bedeutet eine Eins, dass der Austausch ohne Strafe erfolgen kann.



(a) Single-Assay (b) Single-Assay Ausschnitt (c) Multi-Assay (d) Multi-Assay Ausschnitt

Abbildung III.2.5.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Element-Ebene (Knotenlabel „elements“).



(a) Single-Assay 1000 ms (b) Multi-Assay 1000 ms (c) Single-Assay 2000 ms (d) Multi-Assay 2000 ms

Abbildung III.2.6.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypen- und Atomtypengruppen-Ebene (Knotenlabel „fconv“ beziehungsweise „groups“): mit Zeitbegrenzung von 1000 Millisekunden (ms) ((a) und (b)) und mit Zeitbegrenzung von 2000 ms ((c) und (d)).

Je höher der Ähnlichkeitswert, desto ähnlicher sind sich die auszutauschenden Gruppen. Die Gruppen in den Spalten der ersten Reihe, entsprechen denen, die in einem möglichen Hit gefunden werden. Die Ergebnisse sind in Abbildung III.2.7 abgebildet. Die Verwendung der Ähnlichkeitsmatrix 1 führt zu einer leichten Verbesserung auf dem Single-Assay-Datensatz und einer leichten Verschlechterung auf dem Multi-Assay-Datensatz.

Häufig wird einem Atom jedoch nicht nur eine Atomtypengruppe, sondern mehrere zugeordnet. Nach der Ähnlichkeitsmatrix 1 würde jedoch ein Wasserstoffbrückenakzeptor nicht auf einen Wasserstoffbrückendoneptor abgebildet werden, da zwischen ihnen keine Ähnlichkeit in der Matrix definiert wurde. Demzufolge wurde die Ähnlichkeitsmatrix 1 um häufige Kombinationen erweitert und als Ähnlichkeitsmatrix 2 (siehe Tabelle III.2.3) vom BBPMCS-Algorithmus verwendet.

Erstmals kann der MSS-Ansatz, basierend auf fconv-Atomtypengruppen unter Verwendung der Ähnlichkeitsmatrix 2, den MACCS-Key Fingerprint übertreffen (siehe Abbildung III.2.8). Für den Multi-Assay-Datensatz konnte mithilfe der Ähnlichkeitsmatrix 2 das beste Ergebnis unter Verwendung von fconv-Atomtypengruppen erzielt werden.

Gruppe gefunden: / gesucht:	0	1	2	3	4	5	6	7
0	1			0,5	0,6	0,3		
1		1					0,3	
2			1					0,3
3	0,4			1	0,8	0,3		
4	0,6			0,7	1	0,8		
5	0,3			0,3	0,6	1		
6		0,3					1	
7			0,3					1

Tabelle III.2.2.: Ähnlichkeitsmatrix 1 zwischen den Atomtypengruppen. Freie Zellen bedeuten, dass die Atomtypengruppen nicht aufeinander abgebildet werden dürfen.

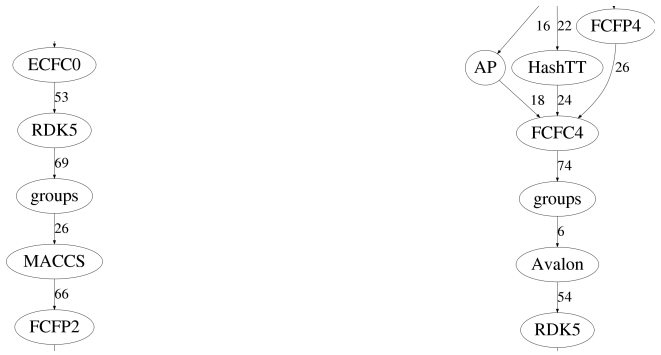


(a) Single-Assay (b) Multi-Assay

Abbildung III.2.7.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypengruppen-Ebene und Ähnlichkeitsmatrix 1 (Knotenlabel „groups“).

Gruppe gefunden: / gesucht:	0	1	2	3	4	5	6	7	40	37	120	10	124	12	123	13	23	237
0	1			0,5	0,6	0,3			1	0,5	1	1						
1		1					0,3				1	1	1	1	1	1		
2			1					0,3			1		1	1	1		1	1
3	0,4			1	0,8	0,3			0,5	1					1	1	1	1
4	0,6			0,7	1	0,8			1		0,3	0,3	1		0,5	0,5	0,5	
5	0,3			0,3	0,6	1												
6		0,3					1											
7			0,3					1		1								1
40	0,8			0,5	0,8				1		0,5	0,5	0,2		0,2	0,2	0,2	
37	0,2			0,8				0,8		1					0,2	0,2	0,2	1
120		0,5	0,5	0,5					0,2	0,2	1	0,5	0,5	0,3	0,8	0,5	0,5	0,3
10	0,8	0,8							0,2	0,2	1	1	0,5	0,3	0,8	0,5	0,2	
124		0,5	0,5		0,5				0,3		0,4	0,3	1	0,5	0,8	0,6	0,6	
12		0,5	0,5								1	0,5	1	1	1	0,5	0,5	0,5
123		0,5	0,5	0,5					0,2	0,3	0,5	0,3	0,3	0,3	1	0,5	0,5	0,5
13		0,8		0,8					0,2	0,5	0,3	0,3	0,3	0,2	1	1	0,5	0,5
23			0,8	0,8					0,2	0,5	0,3		0,3	0,2	1	0,5	1	1
237			0,3	0,4				0,4		0,5	0,2		0,3		0,3		0,5	1

Tabelle III.2.3.: Ähnlichkeitsmatrix 2 zwischen den Atomtypengruppen und einigen Kombinationen von Atomtypengruppen. Freie Zellen bedeuten, dass die Atomtypengruppen bzw. Kombinationen nicht aufeinander abgebildet werden dürfen.

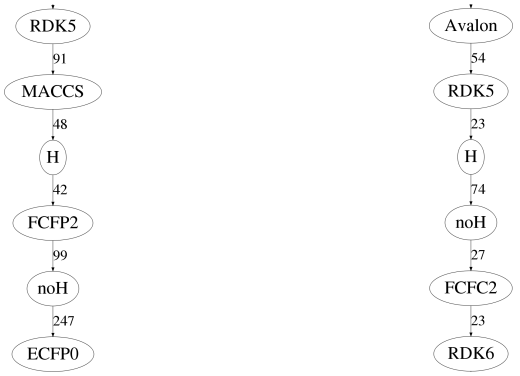


(a) Single Assay (b) Multi Assay

Abbildung III.2.8.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypengruppen-Ebene und Ähnlichkeitsmatrix 2 (Knotenlabel „groups“).

III.2.2.2. Pharmakophor-Ebene

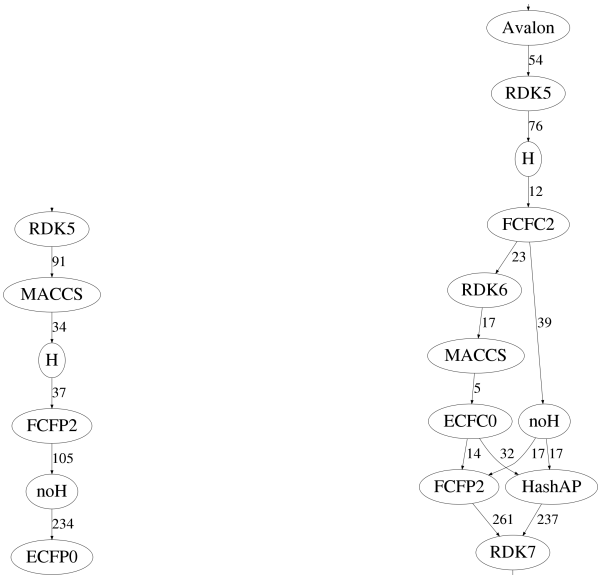
Das Pharmakophorkonzept ist ein häufig angewandtes Konzept in der rationalen Wirkstoffforschung und oft Bestandteil von virtuellen Screening-Workflows (Sanders *et al.* 2012). Somit ist anzunehmen, dass eine Pharmakophor-basierte Ähnlichkeit häufig mit einer Bioähnlichkeit der Moleküle einhergeht. Aus diesem Grund soll in dieser Ebene eine Pharmakophor-basierte Ähnlichkeit analysiert werden. Zur Erstellung eines Pharmakophor-Modells eines Moleküls wurden benachbarte Atome mit gleichem fconv-Atomtyp oder gleicher fconv-Atomtypengruppe mithilfe eines Python-Skriptes von Robert Gehde, das im Rahmen einer Bachelorarbeit entwickelt wurde, zusammengefasst (Gehde 2018). Dabei werden in dem zusammengefassten Label der fconv-Atomtyp bzw. die fconv-Atomtypengruppe und die Anzahl der zusammengefassten Atome gespeichert. Allerdings wird im Rahmen der vorliegenden Arbeit die Anzahl der zusammengefassten Atome nicht verarbeitet und kann in zukünftigen Untersuchungen zur weiteren Verfeinerung des Ergebnisses genutzt werde. Es wurden zwei Varianten getestet. Bei der einen werden die adjazenten Wasserstoffatome beibehalten („H“), während sie in der zweiten Variante entfernt werden („noH“). Eine beispielhafte Veranschaulichung der resultierenden reduzierten Graphen befindet sich in den Abbildungen II.0.2e und II.0.2f. Neben dem Test der beiden Pharmakophorvarianten (siehe Abbildung III.2.9) erfolgte eine Analyse der Kombination dieser mit der Ähnlichkeitsmatrix 2 (siehe Abbildung III.2.10).



(a) Single-Assay (b) Multi-Assay

Abbildung III.2.9.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Pharmakophor-Ebene (Knotenlabel „H“ unter Berücksichtigung von Wasserstoffatomen und „noH“ ohne Berücksichtigung von Wasserstoffatomen).

Die Ergebnisse mit Speicherung der adjazenten Wasserstoffatome zusammengefasster Atome sind in allen vier Varianten (mit und ohne Ähnlichkeitsmatrix sowie Single-Assay und Multi-Assay Datensatz) den Ergebnissen der Varianten ohne diese Wasserstoffatome überlegen.



(a) Single-Assay (b) Multi-Assay

Abbildung III.2.10.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Pharmakophor-Ebene und Ähnlichkeitsmatrix 2 (Knotenlabel „H“ unter Berücksichtigung von Wasserstoffatomen und „noH“ ohne Berücksichtigung von Wasserstoffatomen).

Dies könnte ein Hinweis darauf sein, dass die Anwendung des Pharmakophorkonzeptes, in der hier präsentierten Weise, ohne die Speicherung der adjazenten Wasserstoffatome zusammengefasster Atome, eine zu starke Vereinfachung darstellt, was in einer zu geringen Auflösung für eine zuverlässige Generierung von Aktivitätsrankings resultiert. Die Reduktion der Graphgrößen führt zur Reduktion des Raumes, in den die Scores abgebildet werden können und damit häufiger zu Zusammentreffen, d. h. gleichen Scores von verschieden Molekülen. Um die Möglichkeit einer Abhängigkeit der Performanz auf dem OBDS von der Graphgröße (sowohl vom Eingabe-, d. h. Suchgraphen, als auch vom MSS) zu untersuchen, wurde der Spearman-Rangkorrelationskoeffizient gegen die jeweilige Graphgröße aufgetragen. Dabei ist der Spearman-Rangkorrelationskoeffizient das zur Bewertung der Güte der Ähnlichkeitsbewertung herangezogene Maß und sagt aus, wie gut das Ranking durch die untersuchte Ähnlichkeitsmethode dem der *ground truth* entspricht. Je höher der Spearman-Rangkorrelationskoeffizient ist, desto besser die Übereinstimmung (bei Übereinstimmung der beiden Rankings ist der Spearman-Rangkorrelationskoeffizient eins) und desto besser die verwendete Ähnlichkeitsmethode. Die resultierenden Plots befinden sich im Anhang unter A.4. In beiden Plots ist eine leichte Steigung erkennbar. Der Korrelationskoeffizient (r) beträgt gerundet 0,005 für die Auftragung der Suchgraphengröße (siehe Anhang unter A.3a) und 0,16 für die Auftragung der MSS-Größe (siehe Anhang unter A.3b), der p-Wert ist mit $1,1 \cdot 10^{-31}$ für die Suchgraphengröße bzw. 0 für die MSS-Größe jedoch klein, sodass die Nullhypothese, dass die Steigung Null ist, abgelehnt werden muss und die Korrelation signifikant ist. In Zukunft sollte dieser Effekt genauer untersucht und eine Möglichkeit, den Score unabhängig von der Graphgröße zu berechnen, gesucht werden.

Ein wichtiger Bestandteil des klassischen Pharmakophorkonzeptes ist die Berücksichtigung der dreidimensionalen Anordnung der pharmakophoren Eigenschaften. In der hier vorgestellten Variante des Pharmakophorkonzeptes wird die dreidimensionale Anordnung vernachlässigt. Dies könnte eine weitere mögliche Erklärung für die fehlende Auflösung sein. Ein weiterer Unterschied zum klassischen Pharmakophorkonzept besteht im Umfang der Repräsentation der Moleküle. Während im klassischen Pharmakophorkonzept nur die pharmakophoren Eigenschaften in der Repräsentation abgebildet werden, beinhaltet die Repräsentation in dieser Arbeit zusätzliche Eigenschaften und Atome, welche keiner Pharmakophoreneigenschaft zugeordnet werden können. In der Zukunft könnte die Repräsentation weiter angepasst werden, sodass sie dem klassischen Pharmakophorkonzept weitestgehend entspricht. Interessanterweise ist die Kombination mit der Ähnlichkeitsmatrix 2 für die Pharmakophor-Ebene nur für den Single-Assay-Datensatz - und dort ebenfalls nur für die Variante „H“ - vorteilhaft. Eine mögliche Erklärung ist, dass die Wahrscheinlichkeit, eine ähnliche biologische Aktivität beizubehalten, mit zunehmender Anzahl an Veränderungen sinkt. Wird folglich statt nur einem Atom eine Gruppe ausgetauscht und laut Ähnlichkeitsmatrix 2 bewertet, so ist eine Überschätzung der Ähnlichkeit wahrscheinlich, da die Werte der Ähnlichkeitsmatrix 2 nicht für den Austausch von Gruppen angepasst wurden.

Eine mögliche Lösung bestünde darin, den Wert aus der Ähnlichkeitsmatrix 2 in Abhängigkeit von der Gruppengröße anzupassen. Außerdem könnte es ebenso zu einer Überbewertung der Ähnlichkeit bei dem Vergleich von Gruppen mit gleichem Label jedoch verschiedener Gruppengröße kommen. Auch hier könnte sich die Einbeziehung der Gruppengröße in die Score-Berechnung positiv auswirken.

Zusammengefasst besitzt die vorgestellte Variante des Pharmakophorkonzeptes eine geringere Auflösung bzw. geringere diskriminative Kraft als die Atomtypen- bzw. Atomtypengruppen-Ebene der Ähnlichkeit. Allerdings ist es mit der Pharmakophor-Ebene möglich auch weiter entfernte Ähnlichkeiten zu detektieren und so innovative Strukturen zu finden (Hu, Stumpfe und Bajorath 2017).

III.2.2.3. Analyse verschiedener Ähnlichkeitsfunktionen

Bisher wurden die Experimente mit der Ähnlichkeitsfunktion nach Wallis *et al.* (2001) (Tabelle III.2.4, ID 1) durchgeführt, weil für die Fingerprintbewertung meistens der Tanimoto-Koeffizient (Ausnahmen: AP, TT, alle ECFC- und FCFC-Fingerprints) verwendet wurde und die Ähnlichkeitsfunktion 1 als Tanimoto-Koeffizient, welcher auf Graphen übertragen wurde, verstanden werden kann. Ein kritischer Punkt der vorgestellten Evaluierung von Ähnlichkeitssuchen ist die Bewertung der Ähnlichkeitssuche ausschließlich über die Reihenfolge (dem Ranking) der Ähnlichkeitsscores mit anschließender Bestimmung des Spearman-Rangkorrelationskoeffizienten. Dabei wird die Größe des Abstandes zwischen zwei Ähnlichkeitsscores vernachlässigt. Somit lassen sich unterschiedliche Methoden besser vergleichen. Allerdings wird dieser Vorteil mit einer Vereinfachung erkauft, welche zwei Ähnlichkeitssuchen gleich bewerten kann, obwohl sich das Ausmaß der Fehleinschätzung deutlich unterscheidet. Zum Beispiel ergeben folgende Ergebnisse einer Ähnlichkeitssuche den gleichen Spearman-Rangkorrelationskoeffizienten (0,8):

BBPMCS fconv	0,846154	0,688172	0,725275	0,40909
ECFP4-TC	0,675675675676	0,488888888889	0,634146341463	0,434782608696

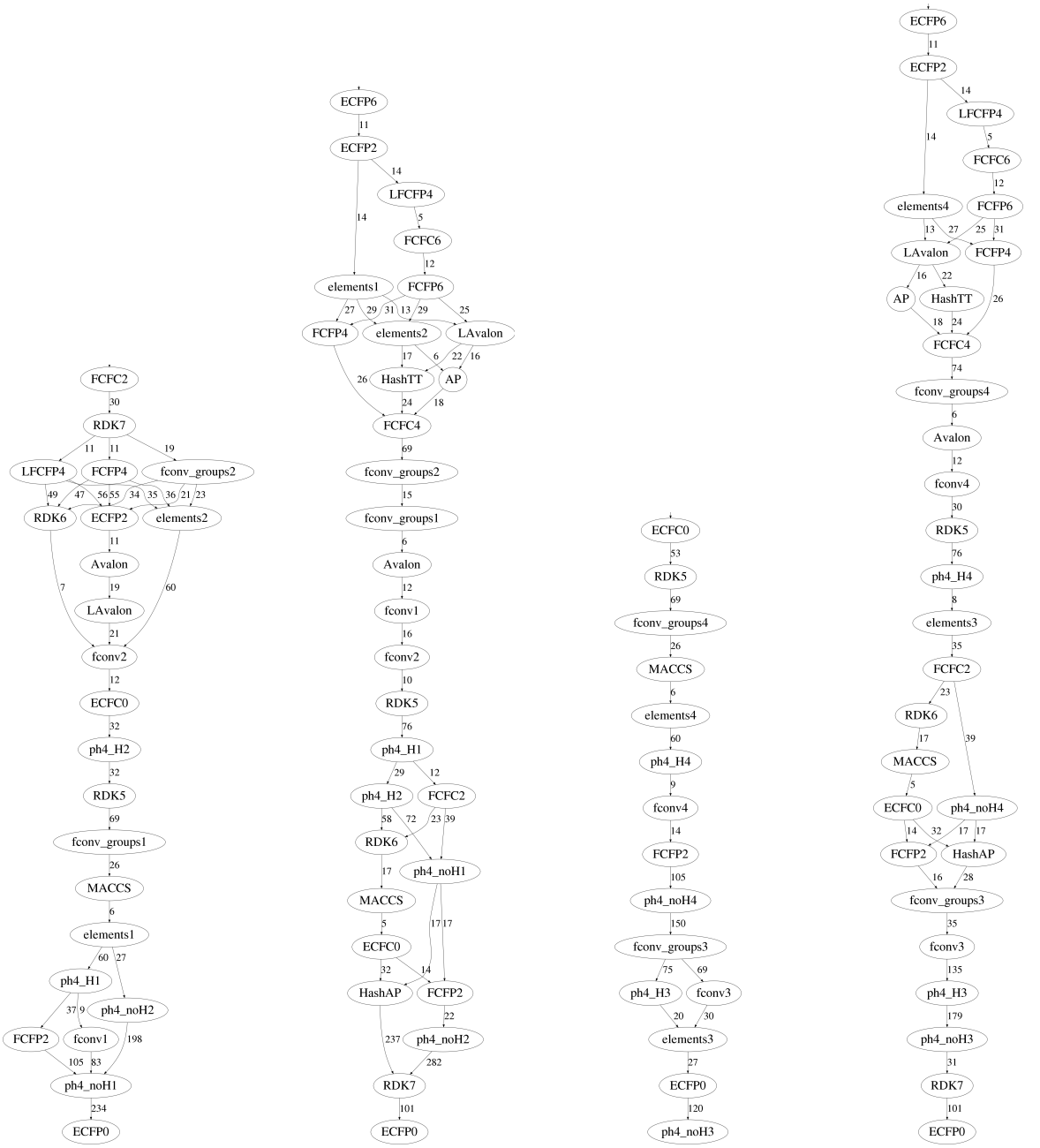
Anzumerken ist, dass demzufolge auch die Anzahl der Nachkommastellen einen möglichen Einfluss besitzt. Jedoch wird durch die geringere Anzahl an Nachkommastellen des BBPMCS-Ergebnisses der MSS-Ansatz tendenziell eher benachteiligt und seine Performanz somit eher unter- und nicht überschätzt. Da neben der verwendeten Ähnlichkeitssuche bzw. Fingerprint-Methode ebenfalls die ausgewählte Ähnlichkeitsfunktion einen Einfluss hat, soll in einem weiteren Experiment der Einfluss der verwendeten Ähnlichkeitsfunktion auf das Ranking und damit die Performanz der BBPMCS Methoden analysiert werden. Dazu wurden insgesamt acht verschiedene Ähnlichkeitsfunktionen, entnommen aus Raymond und Willett (2002), getestet. Sie sind in Tabelle III.2.4 aufgeführt.

Für die Ansätze basierend auf der Pharmakophor- und fconv-Atomtypengruppen-Ebene („ph4_H“, „ph4_noH“ und „fconv_groups“) wurde die Ähnlichkeitsmatrix 2 verwendet. Im Single-Assay-Datensatz ist die Ähnlichkeitsfunktion 2 der Ähnlichkeitsfunktion 1 für alle getesteten Repräsentationen überlegen (siehe Abbildung III.2.11a). Im Multi-Assay-Datensatz fallen die Ergebnisse der Ähnlichkeitsfunktionen 1 und 2 ähnlicher aus (siehe Abbildung III.2.11b). Dort ist die Ähnlichkeitsfunktion 2 nur für den Ansatz unter Verwendung von fconv-Atomtypengruppen der Ähnlichkeitsfunktion 1 überlegen. Der Ansatz unter Verwendung von fconv-Atomtypengruppen ist allen anderen nicht-Fingerprint (BBPMCS) Ansätzen überlegen auf dem Single-Assay-Datensatz für die Ähnlichkeitsfunktionen 1 und 2. Die Ähnlichkeitsfunktion 4 liefert im Allgemeinen bessere Ergebnisse als die Ähnlichkeitsfunktion 3 auf beiden Datensätzen (siehe Abbildungen III.2.11c und III.2.11d). So sind auf dem Single-Assay-Datensatz alle Ansätze unter der Verwendung der Ähnlichkeitsfunktion 4 vor denen unter Verwendung der Ähnlichkeitsfunktion 3. Der Ansatz unter Verwendung von fconv-Atomtypengruppen erzielt für die Ähnlichkeitsfunktionen 3 und 4 auf dem Single-Assay-Datensatz erneut die besten Ergebnisse. Wohingegen dieser Ansatz hinter dem Ansatz unter Verwendung der Element-Label nur das zweitbeste Ergebnis auf dem Multi-Assay-Datensatz für die Ähnlichkeitsfunktionen 1 bis 4 erreicht. In allen Fällen ist unter Verwendung derselben Ähnlichkeitsfunktion der fconv-Atomtypengruppen-Ansatz besser als der fconv-Atomtypen-Ansatz und der Pharmakophor-Ansatz unter Beibehaltung der Wasserstoffatome gegenüber dem ohne Speicherung der Wasserstoffatome bevorzugt.

Für beide Datensätz (Single- und Multi-Assay) ist die Ähnlichkeitsfunktion 6 der Ähnlichkeitsfunktion 5 überlegen (siehe Abbildung III.2.12). Erneut ist für den Single-Assay Datensatz jeweils der Ansatz unter Verwendung von fconv-Atomtypengruppen der am besten bewertete von mir neu entwickelte Ansatz. Auf dem Multi-Assay-Datensatz wird der Ansatz unter Verwendung von Element-Labeln für beide Ähnlichkeitsfunktionen gefolgt vom Ansatz unter Verwendung von fconv-Atomtypengruppen. Für die Ähnlichkeitsfunktionen 7 und 8 ist erkennbar, dass diese einen markanten Einfluss haben, da deutlich weniger signifikante Korrelationen gefunden werden können (siehe Abbildung III.2.13). Die Ähnlichkeitsfunktion, welche die neuen Methoden am besten, im Vergleich zu den Fingerprint-Methoden auf dem Single-Assay-Datensatz bewertet, ist die Ähnlichkeitsfunktion 2. Die Ähnlichkeitsfunktion 3 erzielt die schlechtesten Ergebnisse auf dem Single-Assay-Datensatz. Auf dem Multi-Assay-Datensatz erreichen die Ähnlichkeitsfunktionen 1, 4 und 6 sehr ähnliche Ergebnisse, die jeweils die der Ähnlichkeitsfunktionen 2, 3 und 5 übertreffen. Interessanterweise ist der Ansatz basierend auf fconv-Atomtypengruppen der beste neue Ansatz auf dem Single-Assay-Datensatz, während auf dem Multi-Assay-Datensatz der Ansatz unter Verwendung von Element-Labeln, gefolgt vom fconv-Atomtypengruppen-Ansatz, die besten Ergebnisse liefert. Dies könnte ein Hinweis darauf sein, dass der Ansatz basierend auf fconv-Atomtypengruppen eine genauere Repräsentation erlaubt als der Element-Label basierte Ansatz.

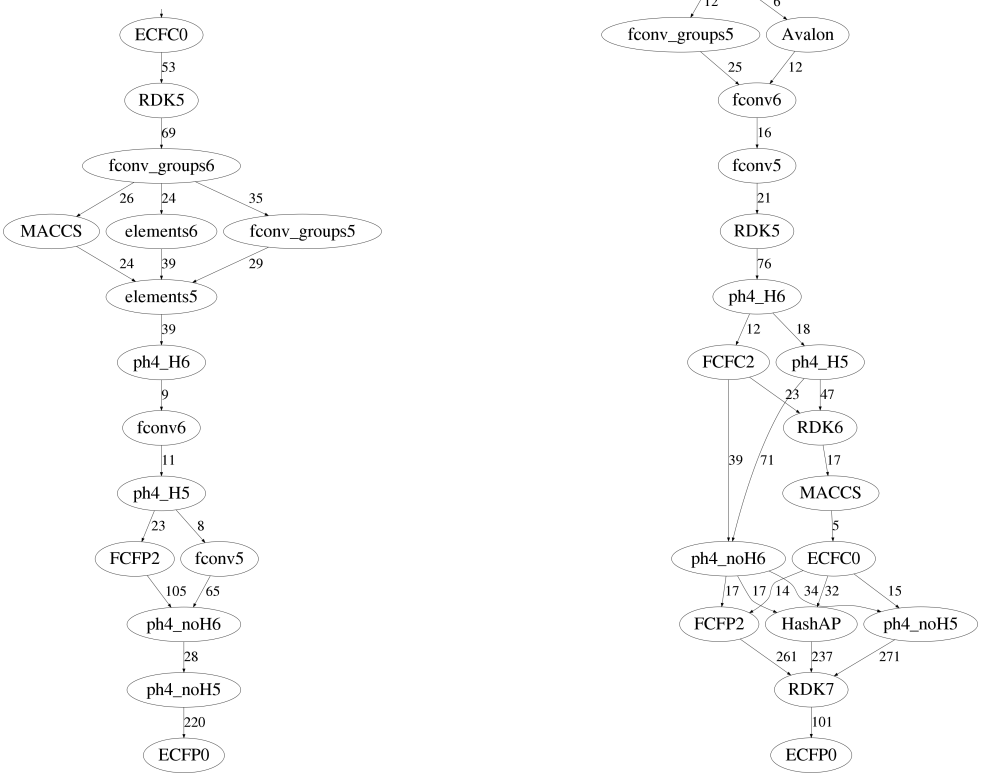
ID	Quelle und ggfs. Name	Gleichung
1	Wallis *et al.* (2001)	$\frac{\lvert G'H'\rvert}{\lvert G\rvert + \lvert H\rvert - \lvert G'H'\rvert}$
2	Bunke und Shearer (1998)	$\frac{\lvert G'H'\rvert}{\max\{\lvert G\rvert, \lvert H\rvert\}}$
3	asymmetrisch oder Szymkiewicz–Simpson (Ellis, Furner-Hines und Willett 1993)	$\frac{\lvert G'H'\rvert}{\min\{\lvert G\rvert, \lvert H\rvert\}}$
4	normalisierter Johnson (Raymond und Willett 2002)	$\frac{2 \cdot \lvert G'H'\rvert}{\lvert G\rvert + \lvert H\rvert}$
5	Johnson (Raymond und Willett 2002)	$\frac{\lvert G'H'\rvert^2}{\lvert G\rvert \cdot \lvert H\rvert}$
6	Sokal und Sneath (Ellis, Furner-Hines und Willett 1993)	$\frac{\lvert G'H'\rvert}{2 \cdot \lvert G\rvert + 2 \cdot \lvert H\rvert - 3 \cdot \lvert G'H'\rvert}$
7	Kulczynski (Ellis, Furner-Hines und Willett 1993)	$\frac{\lvert G'H'\rvert \cdot (\lvert G\rvert + \lvert H\rvert)}{2 \cdot \lvert G\rvert \cdot \lvert H\rvert}$
8	McConnaughey (Ellis, Furner-Hines und Willett 1993)	$\frac{\lvert G\rvert \cdot \lvert G'H'\rvert + \lvert H\rvert \cdot \lvert G'H'\rvert - \lvert G\rvert \cdot \lvert H\rvert}{\lvert G\rvert \cdot \lvert H\rvert}$

Tabelle III.2.4.: Ähnlichkeitsfunktionen die vom BBPMCS-Algorithmus verwendet werden können. Verändert nach Raymond und Willett (2002). In die Berechnungen gehen die Graphgrößen der Graphen G ($|G|$) und H ($|H|$) sowie des MCS $G'H'$ ($|G'H'|$) ein.



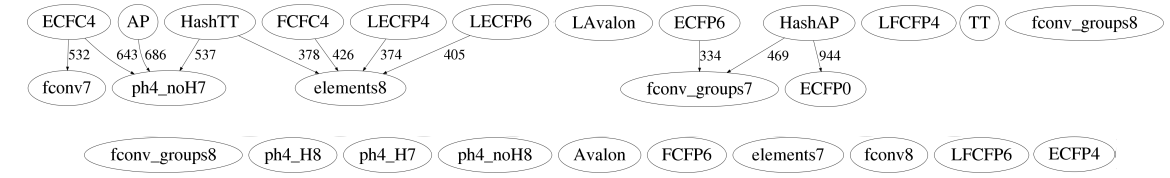
(a) Single-Assay 1 und 2 (b) Multi-Assay 1 und 2 (c) Single-Assay 3 und 4 (d) Multi-Assay 3 und 4

Abbildung III.2.11.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 1 und 2 ((a) und (b)) sowie der Ähnlichkeitsfunktionen 3 und 4 ((c) und (d)) (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen, Zahl: Ähnlichkeitsfunktion).

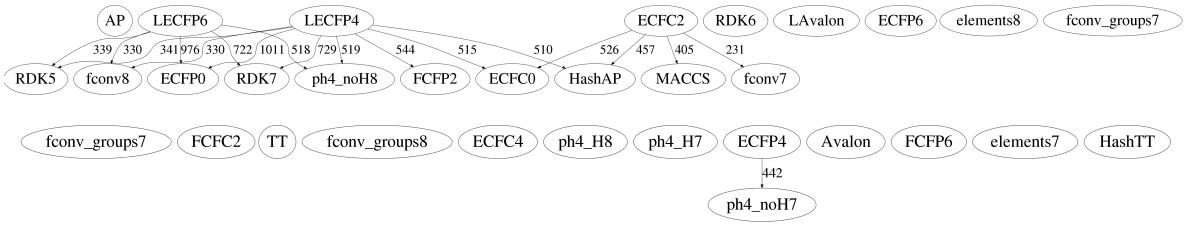


(a) Single-Assay 5 und 6 (b) Multi-Assay 5 und 6

Abbildung III.2.12.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 5 und 6 (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen, Zahl: Ähnlichkeitsfunktion).

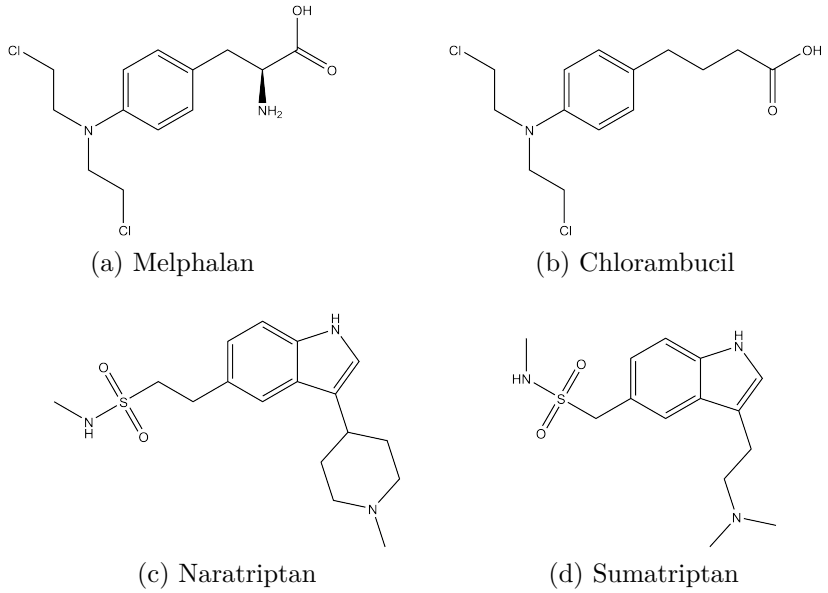


(a) Single-Assay 7 und 8



(b) Multi-Assay 7 und 8

Abbildung III.2.13.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 7 und 8 (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen, Zahl: Ähnlichkeitsfunktion).



(a) Melphalan (b) Chlorambucil

(c) Naratriptan (d) Sumatriptan

Abbildung III.2.14.: Beispiele für ähnliche Moleküle mit Variation der Kettenlänge: die beiden Zytostatika Mephalan (a) und Chlorambucil (b), sowie die beiden Triptane Naratriptan (c) und Sumatriptan (d).

Eine genauere Repräsentation wird zur Detektion von feineren Abstufungen, welche zur Beurteilung von sehr ähnlichen Molekülen, wie denen aus dem Single-Assay-Datensatz, essentiell ist, benötigt. Für weiter entfernte Ähnlichkeiten, wie sie im Multi-Assay-Datensatz simuliert werden sollen, ist eher eine etwas gröbere Abstufung vorteilhaft. Zusammenfassend betrachtet, ist jedoch der MSS-Ansatz zur Detektion und Bewertung weiter entfernter Ähnlichkeiten besser geeignet, weshalb er im Multi-Assay-Datensatz besser als im Single-Assay-Datensatz im Vergleich zu den Fingerprint-Methoden abschneidet.

III.2.2.4. Zulassung von Lücken

Einige Moleküle mit ähnlicher biologischer Aktivität unterscheiden sich durch Kettenlängen oder allgemeiner durch die Anzahl an Atomen, welche zwei identische Bereiche verbinden (Beispiele siehe Abbildung III.2.14). Um auch diese Ähnlichkeit adäquat abbilden zu können, wurde der Suchalgorithmus (BBPMCS) erweitert (Droschinsky, Kriege und Mutzel 2018). Dazu wurde der *p-Parameter* (*gap penalty*) eingeführt, welcher die Höhe der Strafe für jedes ausgelassene Atom angibt (Gap-Strafe). Die Größe des MSS ($|G'H'|$) wird demzufolge um das Produkt aus Gap-Strafe (p-Parameter) und Anzahl ausgelassener Atome reduziert. Es wurden Gap-Strafen von: 0, 1, 2, 3, 5, 8 und 13 getestet.
Die Ergebnisse, welche unter Verwendung der Ähnlichkeitsmatrix 2 und der Ähnlichkeitsfunktion 2 erhalten wurden, sind in den Abbildungen III.2.15 und III.2.16 dargestellt.

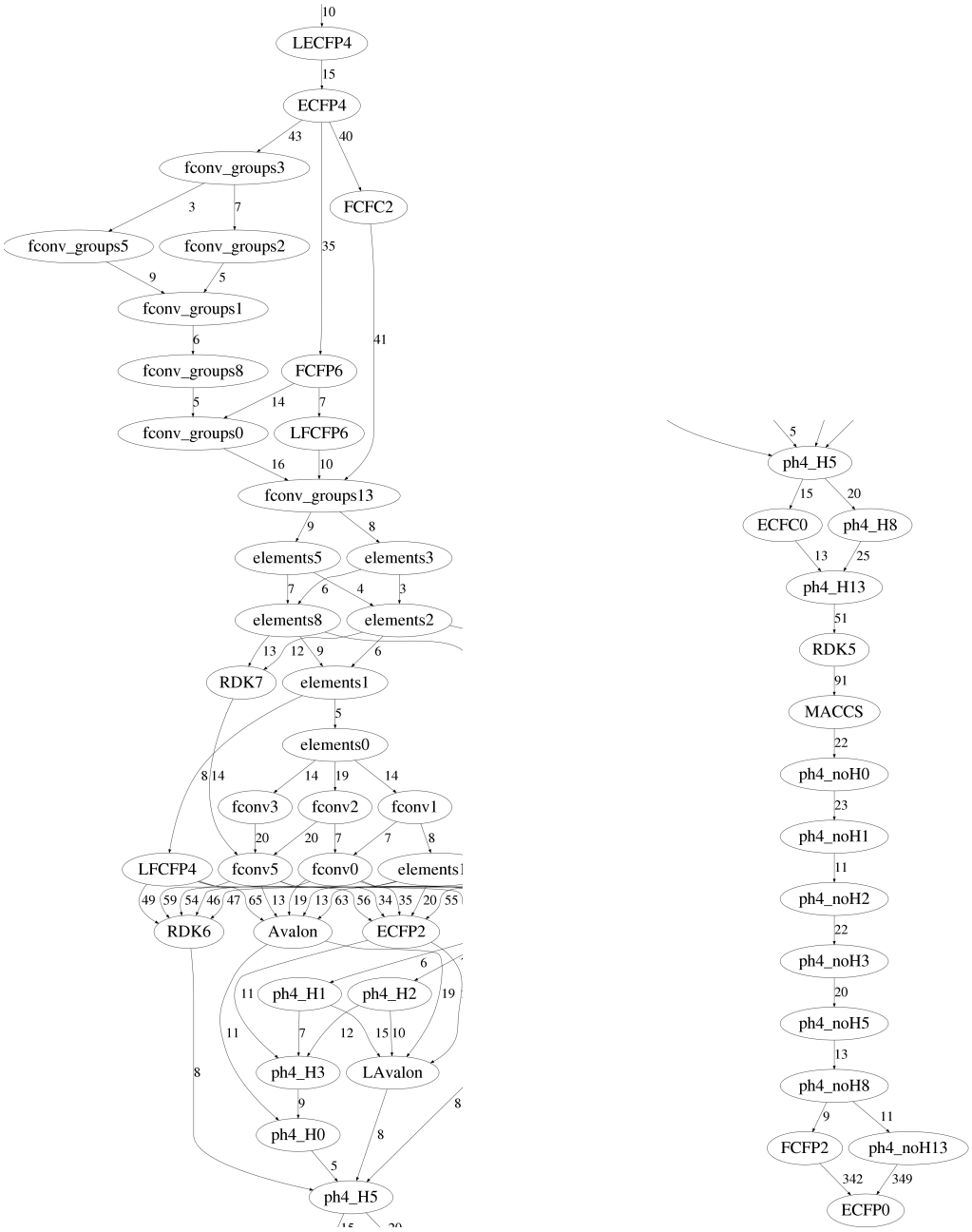


Abbildung III.2.15.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Zulassung von Lücken für den Single-Assay Benchmarkdatensatz. Die Zahl hinter der Abkürzung der Methode gibt die verwendete Gap-Strafe an (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen).

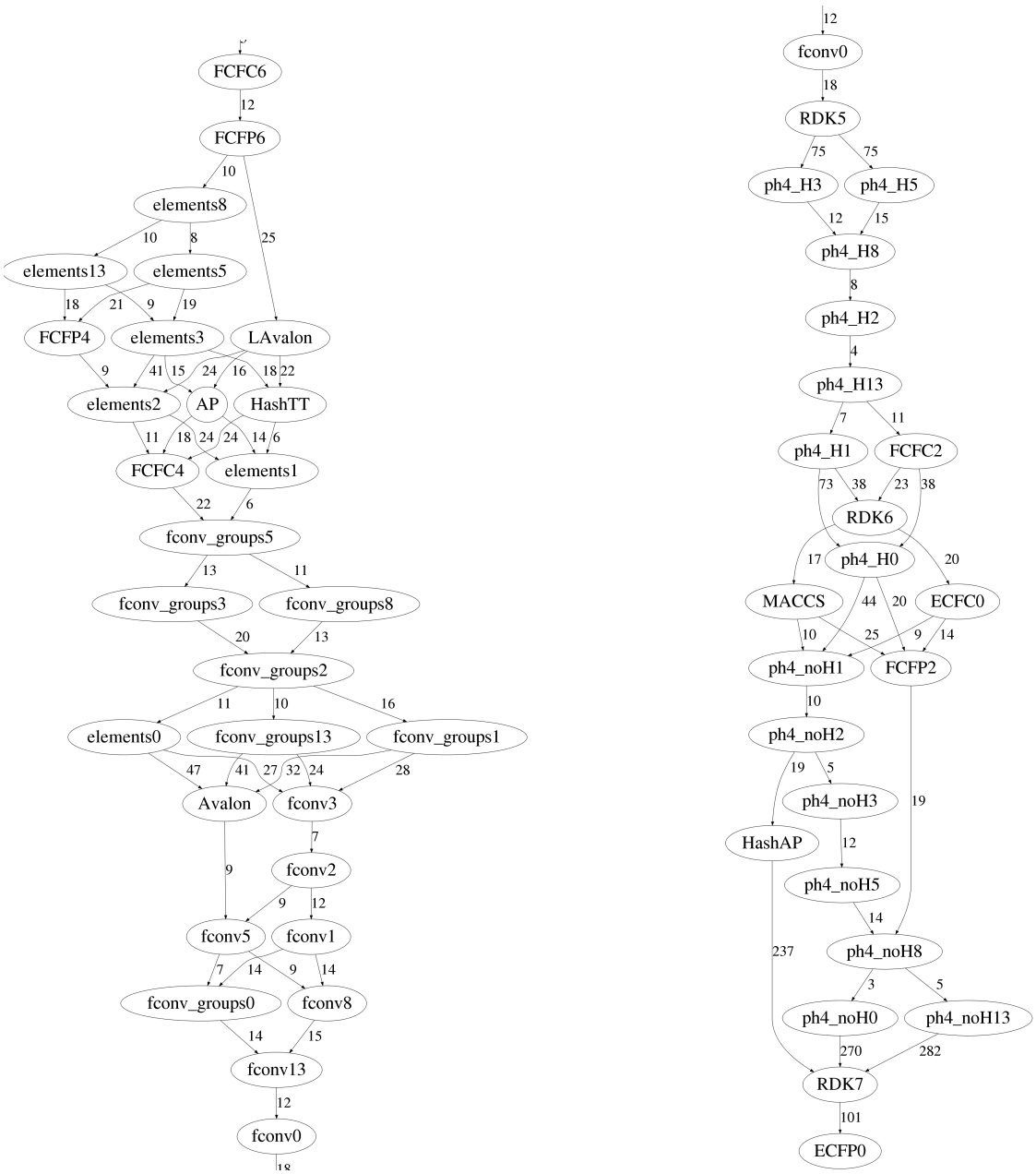


Abbildung III.2.16.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Zulassung von Lücken für den Multi-Assay Benchmarkdatensatz. Die Zahl hinter der Abkürzung der Methode gibt die verwendete Gap-Strafe an (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen).

Auf dem Single-Assay-Benchmarkdatensatz kann mit einer Gap-Strafe von 3 und unter der Verwendung von fconv-Atomtypengruppen das bisher beste Ergebnis erzielt werden. Eine Gap-Strafe von 3 liefert außerdem für die anderen nicht Pharmakophor-basierten Ansätze (mit fconv-Atomtypen und Elementlabeln) bessere Ergebnisse als ohne Zulassung von Lücken. Auch der Pharmakophor-basierte Ansatz unter Beibehaltung der Wasserstoffatome profitiert von der Zulassung von Lücken. Interessant ist, dass sich die Ergebnisse des Pharmakophor-basierten Ansatzes ohne Beibehaltung der Wasserstoffatome mit steigender Gap-Strafe verschlechtern.

Die Auswirkungen einer Zulassung von Lücken auf die Ergebnisse vom Multi-Assay-Datensatz sind weitestgehend vernachlässigbar. Hier kristallisiert sich keine präferierte Gap-Strafe heraus. Wobei das Ergebnis für eine Gap-Strafe von 3 für alle getesteten Ansätze immer unter den besten drei Ergebnissen des jeweiligen Ansatzes zu finden ist. Daraus kann geschlussfolgert werden, dass die Zulassung von Lücken im Allgemeinen zu einer Verbesserung oder nur geringen Veränderung der Ergebnisse führt. Für einen ersten Test kann demzufolge eine Gap-Strafe von 3 empfohlen werden.

III.2.2.5. Clusteringergebnisse von StruClus unter Verwendung von verschiedenen Repräsentationen und Ähnlichkeitsebenen für den SBDS

Es sollte ebenfalls der Nutzen verschiedener Repräsentationen und Ähnlichkeitsebenen für das Clustering mittels StruClus analysiert werden. Zusätzlich wurden die Ergebnisse von StruClus mit denen eines in der Chemieinformatik häufig verwendeten Clustering-Algorithmus (Jarvis-Patrick-Clustering) verglichen. Zu erwähnen ist, dass: (i) StruClus bereits inheränt selbstständig Ähnlichkeiten identifizieren kann, in dem mehrere Repräsentanten für einen Cluster gewählt werden können, (ii) in der vorliegenden Version von StruClus eine Verwendung von Ähnlichkeitsmatrizen nicht vorgesehen ist und (iii) die Pharmakophorebene mit einer Repräsentation, welche die Wasserstoffatome beibehält, Graphen generiert, welche aktuell nicht mit StruClus verarbeitet werden können. Die Ergebnisse sind in der Tabelle III.2.5 zusammengefasst. Detaillierte Plots sowie ungewichtete Reinheitswerte (siehe Abschnitt IV.2.8.1 und Gleichung IV.2.8) befinden sich im Anhang unter A.9. An den ungewichteten Reinheitswerten des randomisierten Clusterings kann sehr gut erkannt werden, dass das Jarvis-Patrick-Clustering bereits eine erhöhte ungewichtete Reinheit im randomisierten Clustering besitzt. Dies liegt wahrscheinlich an dem hohen Anteil an Singelton-Clustern. Daraus wird ebenfalls ersichtlich, dass die ungewichtete Reinheit nicht als alleiniges Maß für die Qualität eines Clusterings geeignet ist. Zum Beispiel wäre die Reinheit maximal, wenn jedes Molekül alleine einem eigenen (Singleton) Cluster zugeordnet werden würde. Aus diesem Grund wurde einerseits die Reinheit gewichtet (siehe Abschnitt IV.2.8.1 und Gleichung IV.2.10) und andererseits zusätzlich der F1-Score (siehe Abschnitt IV.2.8.1 und Gleichung IV.2.5) als Qualitätskriterium verwendet.

Sowohl für die Reinheit als auch für den F1-Score liegt der Wertebereich zwischen null und eins und ein höherer Wert zeigt in Bezug auf diesen Parameter ein besseres Clusteringergebnis an.

Diese Analyse zeigt sehr deutlich, dass die am besten geeignete Repräsentation und Ähnlichkeitsebene vom betrachteten Target-Protein abhängt. Die Sybyl-Atomtypen-Repräsentationen sowie die Pharmakophor-Ebene konnten für kein Protein das beste Ergebnis erzielen. Die fconv-Atomtypengruppen-Ebene ergibt in zwei Fällen (Acetylcholinesterase (AChE) und COX-2) den zweitbesten gewichteten Gesamt-Reinheitswert und für die *cyclin-dependent kinase 2* (Cyklin-abhängige Kinase 2) (CDK2) den besten gewichten Gesamt-Reinheitswert. Da in diesem Versuch noch keine Kombination von StruClus mit einer Ähnlichkeitsmatrix für die fconv-Atomtypengruppen- und Pharmakophor-Ebene verwendete wurde, ist zu erwarten, dass sich das Ergebnis dieser beiden Ebenen durch die Verwendung einer Ähnlichkeitsmatrix weiter verbessern wird. StruClus ist für die vier Proteintargets AChE, Dopamin D4-Rezeptor, Matrix-Metalloproteinase 3 und Thrombin das beste Verfahren. Das Jarvis-Patrick-Clustering liefert nur für den Angiotensin-II-Rezeptor den besten gewichteten Gesamt-Reinheitswert. Die fconv-Atomtypen-Repräsentation kann für die Proteine Estrogenrezeptor und COX-2 den höchsten gewichteten Gesamt-Reinheitswert erzielen. Unabhängig vom Target betrachtet, konnte das Jarvis-Patrick-Clustering den besten gewichteten F1-Score ($\sim$ 0,47) erreichen, danach folgen StruClus ($\sim$ 0,44) und die fconv-Atomtypengruppen-Ebene ($\sim$ 0,40). Beim ungewichteten F1-Score erzielten alle MSS-Verfahren mit StruClus ähnlich gute Ergebnisse. Das original StruClus-Verfahren erzielte jedoch den niedrigsten ungewichteten F1-Score ($\sim$ 0,12).

Target-Protein	Jarvis-Patrick	original StruClus	Elemente	fconv	fconv-Gruppen	Sybyl	ph4_noH
AChE	0,0020	0,0092	0,0020	0,0009	0,0026	0,0009	0,0023
	0,0007	0,0008	0,0005	0,0006	0,0011	0,0001	0,0006
D4	0,0087	0,0160	0,0126	0,0014	0,0013	0,0008	0,0096
	0,0002	0,0046	0,0008	0,0004	0,0026	0,0004	0,0006
An2	0,0107	0,0017	0,0056	0,0022	0,0009	0,0025	0,0021
	0,0003	0,0018	0,0009	0,0001	0,0026	0,0007	0,0004
ER	0,0043	0,0055	0,0006	0,0060	0,0004	0,0023	0,0032
	0,0004	0,0019	0,0003	0,0007	0,0020	0,0001	0,0003
MMP	0,0038	0,0063	0,0006	0,0012	0,0024	0,0001	0,0013
	0,0004	0,0022	0,0010	0,0003	0,0052	0,0003	0,0007
H3	0,0079	0,0027	0,0105	0,0027	0,0066	0,0019	0,0009
	0,0004	0,0025	0,0010	0,0008	0,0008	0,0005	0,0004
Thrombin	0,0034	0,0044	0,0015	0,0030	0,0009	0,0003	0,0006
	0,0003	0,0023	0,0004	0,0010	0,0010	0,0001	0,0006
COX-2	0,0010	0,0033	0,0025	0,0058	0,0049	0,0002	0,0030
	0,0006	0,0015	0,0013	0,0001	0,0029	0,0005	0,0005
CDK2	0,0013	0,0037	0,0022	0,0010	0,0048	0,0012	0,0008
	0,0005	0,0030	0,0005	0,0011	0,0019	0,0001	0,0001
gewichteter F1-Score	0,4741	0,4392	0,3078	0,3529	0,4015	0,3143	0,3327
	0,1583	0,2828	0,1921	0,1835	0,2422	0,1610	0,1370
F1-Score	0,1736	0,1226	0,2105	0,1958	0,1934	0,1801	0,1990
	0,0852	0,0758	0,1408	0,1055	0,0893	0,0874	0,1121

Tabelle III.2.5.: Ergebnisse von StruClus unter Verwendung von verschiedenen Repräsentationen und Ähnlichkeitsebenen für den SBDS. Angegeben ist die gewichtete Gesamt-Reinheit bzw. der F1-Score (gewichtet und ungewichtet) für das Experiment (obere Reihe) und ein randomisiertes Clustering (untere Reihe). Proteinabkürzungen: Acetylcholinesterase (AChE), Angiotensin-II-Rezeptor (An2), Cyclin-abhängige Kinase 2 (CDK2), Cyclooxygenase-2 (COX-2), Dopamin D4-Rezeptor (D4), Estrogenrezeptor (ER), Histamin-H3-Rezeptor (H3), Matrix-Metalloproteinase 3 (MMP). Grün: höchster Wert, lila: niedrigster Wert. ph4_noH: Pharmakophor-Ebene ohne Berücksichtigung von Wasserstoffatomen.

III.2.3. Zusammenfassung

Eine neue Methode zur Identifizierung und Bewertung von Ähnlichkeit wurde entwickelt. Dieser MSS-Ansatz verknüpft einen Graph-basierten Ähnlichkeitsalgorithmus (Suche oder Clustering) mit einer biologisch-chemischen Vorgabe zu ähnlichen Knotenlabeln. In einer Proof of Concept Studie konnte der Nutzen der innovativen MSS-Methode gezeigt werden und im Zusammenhang mit der anschließenden großangelegten Benchmark-Analyse zeichnet sich eindrucksvoll ab, dass der MSS-Ansatz seine volle Stärke erst auf großen Datensätzen (Big Data) entfaltet und dort wesentlich differenziertere Ergebnisse liefert als der State of the Art Ansatz (ECFP4-TC). Dennoch konnte der MSS-Ansatz bereits im Benchmarktest Ergebnisse erzielen, die mit denen von bekannten Fingerprint-Methoden vergleichbar sind.

Des Weiteren wurden verschiedene Varianten des MSS-Ansatzes in der Benchmark-Analyse mit den bestehenden Fingerprint-Methoden verglichen und erzielten ähnlich gute Ergebnisse. Für die Pharmakophor-Ebene zeichnet sich jedoch ab, dass die in dieser Arbeit verwendete Umsetzung eine zu starke Vereinfachung der Molekülrepräsentation ergibt. Im Gegensatz dazu lieferte die Atomtypengruppen-Ebene gute Ergebnisse. Zusätzlich konnte gezeigt werden, dass der MSS-Ansatz besser für die Detektion und Bewertung entfernter Ähnlichkeiten geeignet ist, als für die von Ähnlichkeiten von sehr nahe verwandten Molekülen. Für einen ersten Test konnten die folgenden Einstellungen als gute Startwerte ermittelt werden: (i) Verwendung der Ähnlichkeitsmatrix 2, (ii) Verwendung der Ähnlichkeitsfunktion 2 und (iii) Gap-Strafe 3. Eine wichtige Lehre aus diesem Kapitel ist, dass eine Balance zwischen chemisch-molekularer Genauigkeit und biologischer *Fuzziness* gewahrt werden muss. Letztendlich ist es jedoch wahrscheinlich, dass diese biologische *Fuzziness* final zu einer biologischen Genauigkeit verfeinert werden muss. Es scheint, als würde sich dieser finale Schritt der aktuellen Betrachtung entziehen, da er losgelöst von der molekularen Struktur und den allgemeinen Regeln zur biologischen Aktivität ist und maßgeblich durch das Zielprotein und dessen Erkennungsstrukturen von Ähnlichkeit beeinflusst wird. Dennoch wurde bereits in diesem Kapitel die biologische *Fuzziness* eingegrenzt und kontrolliert durch die rationale Definition von ähnlichen Atomtypen. Folglich ist die Möglichkeit, die MSS-Suche darüberhinaus gezielt an das Zielprotein über die Ähnlichkeitsmatrix oder die Definition von eigenen Atomtypen bzw. generell Knotenlabeln anzupassen, nicht nur ein Alleinstellungsmerkmal gegenüber den Fingerprint-Methoden, sondern sie wird maßgeblich zum Erfolg des MSS-Ansatzes beitragen. Das Clustering des SBDS zeigt, dass die Performanz eines Ansatzes stark vom untersuchten Protein abhängt. Aufgrund der Wichtigkeit der Auseinandersetzung mit dem Zielprotein bei der Betrachtung der Bioähnlichkeit wird im folgenden Kapitel die Scaffold-Ebene unter Berücksichtigung der Proteinzielstruktur (Protein-Ebene) betrachtet. In Zukunft wird die Nachfrage nach effizienten und verlässlichen Ähnlichkeitssuchen und Clusteringverfahren auf großen Datensätzen durch die stetig steigende Anzahl synthetisierter und auf biologischen Zielstrukturen getesteter Verbindungen weiter ansteigen.

Mit dem MSS-Ansatz wurde dafür ein vielversprechendes Verfahren entwickelt. Als abschließende Bemerkung sei darauf hingewiesen, dass der Nutzen des MSS-Ansatzes über die Erfassung möglicher ähnlicher Bioaktivitäten hinausgeht. Zum Beispiel kann der MSS-Ansatz ebenfalls zur Vorhersage von potentiellen Nebenwirkungen, toxikologischen Effekten oder möglichen Einflüssen auf die Umwelt genutzt werden.

Quintessenz:

- Etablierung des innovativen MSS-Ansatzes für Ähnlichkeitsanalysen und Identifizierung geeigneter Startparameter
- Vorteile des neuen MSS-Ansatzes konnten in einer Proof of Concept Studie gezeigt werden
- in einer großangelegten Benchmarkstudie konnte die Konkurrenzfähigkeit des MSS-Ansatzes in Bezug auf Fingerprint-Methoden demonstriert werden
- die Einbeziehung des Zielproteins in die Betrachtung der Ähnlichkeit ist von essentieller Bedeutung

III.3. Die Scaffold- und Protein-Ebene der Ähnlichkeit

III.3.1. Einleitung

Die in dem vorherigen Kapitel dargestellten Untersuchungen zu den Ebenen der Ähnlichkeit niedermolekularer Verbindungen zeigten, dass auch die Proteine als Interaktionspartner einen wesentlichen Einfluss haben. Aus diesem Grund sollte nun ebenfalls die Zielstruktur in die Ähnlichkeitsbetrachtung mit einbezogen werden. Aus dieser Betrachtung ergibt sich dann die Protein-Ebene der Ähnlichkeit. Die Einbeziehung der Proteine in die Ähnlichkeitsbetrachtung sollte anhand der Verknüpfung der Scaffold-Ebene mit der Protein-Ebene realisiert werden. In der Scaffold-Ebene werden Grundgerüste (Scaffolds) betrachtet, welche Teil (Substruktur) distinkter Liganden für verschiedene Proteine sind (*privilegierte Scaffolds*).

Eine leitende Frage der in diesem Kapitel behandelten Arbeiten war, ob eine Ähnlichkeit der Liganden (basierend auf einem gleichen - privilegierten - Scaffold) auf eine Ähnlichkeit der Proteine, welche durch Liganden mit dieser Ähnlichkeit (demselben Scaffold) moduliert werden, übertragen werden kann. Besitzen Proteine, die durch verschiedene Liganden, mit demselben privilegiertem Scaffold, moduliert werden können, also eine Ähnlichkeit, welche sich nicht aufgrund einer hohen Sequenzähnlichkeit oder einer hohen funktionellen Ähnlichkeit zeigt? Könnte man zeigen, dass die Frage zumindest in bestimmten Fällen positiv zu beantworten ist, so wäre dies die Basis für eine innovative Wirkstoffdesign-Strategie, die sowohl eine frühe Identifizierung möglicher *off-targets* erlaubt, als auch potentielle Kombinations-Ansätze (*Polypharmakologie*) aufzeigt. Kombinations-Ansätze könnten die Effizienz durch gezielte Beeinflussung mehrerer definierter Zielstrukturen steigern. Somit würde diese neuartige Strategie einige Risiken, welche häufig für das Fehlschlagen von Arzneistoffkandidaten in klinischen Studien verantwortlich sind (zu geringe Wirksamkeit oder zu hohe Nebenwirkungen (Harrison 2016)), minimieren. Eine solche Strategie wurde in dieser Arbeit entwickelt und *APSC* (*advanced privileged scaffold concept*) genannt, da nicht nur eine privilegierte Struktur im Liganden sondern ebenfalls im Zielprotein mit einbezogen wird.

Zunächst wurden erweiterte privilegierte Scaffolds und ihre Targets mittels des in dieser Arbeit entwickelten SPARROW-Workflows (*Workflow Of Recognition of Reasonable Advanced Privileged Scaffolds*) aus der DrugBank extrahiert. Anschließend erfolgte eine genaue Analyse der Zusammenhänge der mit SPARROW gefundenen fünf Targets sowie deren Liganden und einem erweiterten privilegierten Scaffold (Chromon). Dazu wurden ähnliche Moleküle zu bekannten Liganden von vier der Targets auf ihre Aktivität bezüglich des fünften Proteins (BRD4) hin untersucht. Parallel wurde zur Kontrolle ein Docking-basierter Ansatz verfolgt. Die hochgescorten Docking-Vorschläge und die APSC-Vorschläge wurden in einem initialen Screen per HTRF-Assay analysiert. Von ausgewählten Hits wurden IC_{50}-Werte bestimmt und orthogonale Assays (MST und TSA) durchgeführt. Für fünf Inhibitoren konnte im Rahmen dieser Arbeit darüberhinaus ein Kokristall mit BRD4 erhalten werden. Eine Literaturrecherche ergab, dass die gefundenen Ähnlichkeiten der Proteine eine physiologische funktionelle Beziehung reflektieren, welche für Strategien zur Bekämpfung von Resistenzen ausgenutzt werden kann.

III.3.1.1. Privilegierte Scaffolds

Eine Strategie, um den Wirkstoffentwicklungsprozess zu beschleunigen, ist die Verwendung von privilegierten Grundgerüsten (Scaffolds). Ziel ist es, dadurch die Wahrscheinlichkeit, einen Hit zu finden, zu erhöhen. Die Annahme ist, dass privilegierte Scaffolds eine Basis für die benötigte dreidimensionale Anordnung der Selektivität-bringenden Substituenten zur Verfügung stellen.

Die ursprüngliche Definition von privilegierten Scaffolds durch Evans *et al.* (1988) wurde immer wieder erweitert. Zum einen sind Scaffolds interessant, die ausschließlich vermehrt in Liganden innerhalb einer Proteinfamilie auftreten (*selective privileged scaffolds*) und zum anderen solche, welche als Substruktur in Liganden verschiedener Proteine vorkommen. Eine Übersicht von bekannten privilegierten Scaffolds befindet sich in Welsch, Snyder und Stockwell (2010). Interessant ist, dass in Naturstoffen häufig privilegierte Scaffolds gefunden werden können. In dieser Arbeit werden Grundgerüste von Molekülen, welche an verschiedene Proteine (idealerweise aus verschiedenen Proteinklassen) binden, privilegierte Scaffolds genannt. In einer anderen Sichtweise könnten diese Scaffolds auch als *promiscious scaffolds* (promuskuitive Scaffolds) deklariert werden. Tatsächlich finden Analysen zu privilegierten Scaffolds wie in Welsch, Snyder und Stockwell (2010) und promiskuitiven Liganden z. B. von Sturm *et al.* (2012) ähnliche Scaffolds. Allerdings kann ein gefundener Scaffold-basierter Hit hinsichtlich seiner Selektivität durch die Wahl der Substituenten optimiert werden. Somit ist Promiskuität weniger eine Eigenschaft des Scaffolds als viel mehr eine Eigenschaft eines Moleküls bzw. Liganden (Schneider und Schneider 2017). Zudem kann mit der gewählten Definition überprüft werden, ob ansonsten unabhängige Proteine Ähnlichkeiten aufweisen, welche die Bindung desselben Scaffolds ermöglichen.

Diese beiden Faktoren zusammen sind wichtig, da zur Erzielung von Selektivität die Kenntnis über Ähnlichkeiten der Proteine sehr nützlich sein kann. Diese Ähnlichkeit könnte in Form von privilegierten Strukturen in verschiedenen Proteinen, welche die Bindung von verschiedenen Liganden, jedoch mit demselben Scaffold, ermöglichen, auftreten. Solche privilegierten Strukturen von Proteinen werden im nächsten Absatz eingeführt.

III.3.1.2. Privilegierte Strukturen in Proteinen und das APSC

Privilegierte Strukturen in im Hinblick auf Sequenz und Funktion verschiedenen Proteinen sind Bereiche, welche ähnliche Eigenschaften, wie eine ähnliche räumliche Anordnung von Sekundärstrukturelementen oder von physikochemischen Eigenschaften, aufweisen und verschiedene Liganden binden. Besitzen die distinkten Liganden dasselbe Scaffold, kann das APSC angewendet werden. Das APSC basiert auf einem Ansatz von Breinbauer, Vetter und Waldmann (2002), welcher eine Einteilung der Proteine auf der Grundlage von ähnlichen Sekundärstrukturen der Bindetaschen vorsieht. Das Ziel beider Ansätze ist die Identifizierung von neuen Modulatoren mithilfe von bekannten Modulatoren ähnlicher Proteine. Zudem können beide Ansätze in der Wirkstoffentwicklung genutzt werden, um frühzeitig mögliche *off-Targets* zu identifizieren oder bewusst mehrere Proteine für eine bessere Wirkung zu adressieren (*Polypharmakologie*). Die Grundannahmen beider Ansätze sind ebenfalls ähnlich. So wird angenommen, dass die Proteinstruktur konservierter ist als die Sequenz, weshalb basierend auf der Struktur noch Ähnlichkeiten detektiert werden können, die durch die Evolution in der Sequenz verborgen sind und dass es trotz einer relativ hohen Anzahl an Proteinen eine vergleichsweise geringe Anzahl möglicher Faltungsmuster gibt, was bedeutet, dass zwischen vielen Proteinen Ähnlichkeiten in der Faltung existieren (Koch *et al.* 2004). Die Ansätze unterscheiden sich jedoch in ihrem Ausgangspunkt. Der Ansatz von Breinbauer, Vetter und Waldmann (2002) startet mit bekannten Proteinstrukturen oder Sequenzen, wohingegen das in dieser Arbeit entwickelte APSC erst basierend auf bekannten Liganden Ähnlichkeiten zwischen Proteinen definiert und somit auf Bioaktivitätsdaten fußt. Zudem ist das APSC nicht auf eine ähnliche räumliche Anordnung von Sekundärstrukturelementen beschränkt.

III.3.2. Ergebnisse und Diskussion

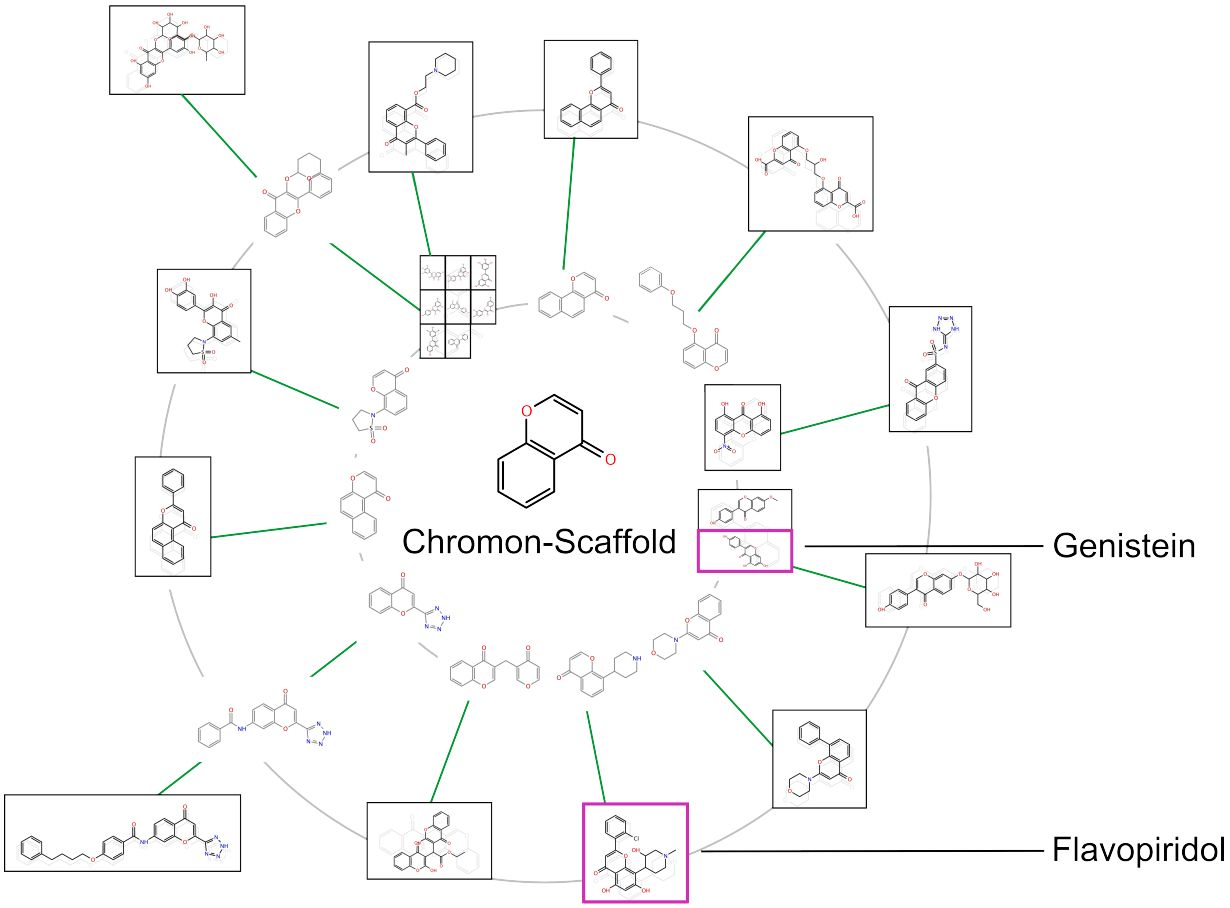
III.3.2.1. SPARROW-Analyse der DrugBank

Zur Untersuchung von Proteinen, welche gleiche Scaffolds als Bestandteil unterschiedlicher Liganden binden, jedoch verschieden in der Gesamtfaltung und der Funktion sind, wurde ein Python-Skript entwickelt. Das Skript wurde SPARROW (*Workflow Of Recognition of Reasonable Advanced Privileged Scaffolds*) genannt. Aus den Liganden werden die Scaffolds mittels einer Kommandozeilen-Version von Scaffold Hunter (Schäfer *et al.* 2017) extrahiert.

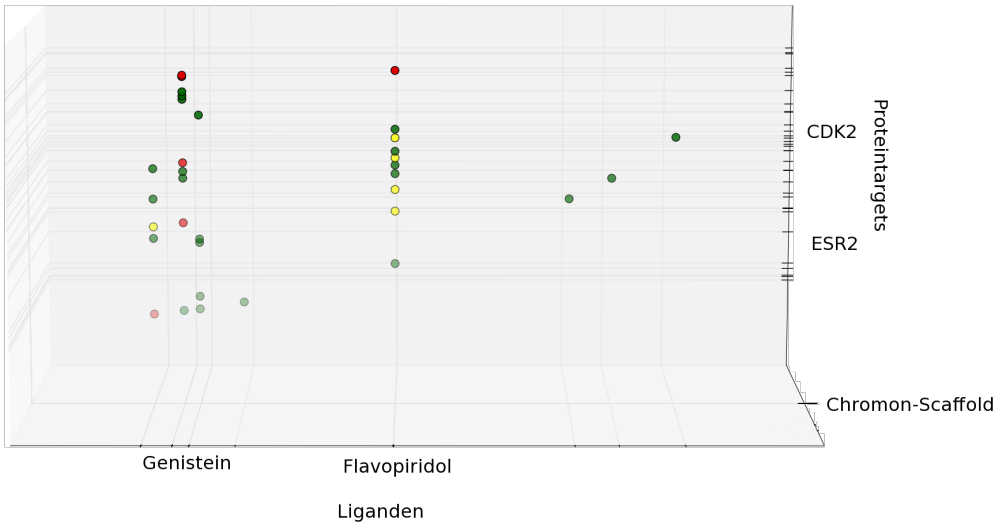
Scaffold Hunter reduziert das Molekül schrittweise auf die zugrundeliegenden Scaffolds und generiert so den Scaffold Baum. Dabei werden in der ersten Ebene z. B. die azyklischen, terminalen Substituenten entfernt. Danach erfolgt in jeder Ebene nach chemischen und medizinisch-chemischen Regeln eine Reduktion um einen Ring, sodass weniger charakteristische Ringe zuerst entfernt werden (Wetzel *et al.* 2009; Schuffenhauer *et al.* 2007). In dieser Arbeit wurde die Ebene der Zwei-Ring-Scaffolds betrachtet. Ein Beispiel für die Scaffold-Ebenen basierend auf Scaffold Hunter ist in der Abbildung III.3.1a gezeigt. Die funktionelle Ähnlichkeit wurde anhand der EC-Nummer bestimmt und die Sequenzähnlichkeit mittels NCBI BLAST+ (Camacho *et al.* 2009) ermittelt. Proteine, welche dasselbe Scaffold binden, jedoch einen Bit-Score von größer 200 und/oder ein e-Value von kleiner als 0,001 bei der BLAST+ Analyse erhielten, wurden als unerwünscht markiert. Zudem wurde überprüft, ob zu den Proteinen eine verfügbare Kristallstruktur in der PDB (Berman *et al.* 2000; *RCSB PDB - Research Collaboratory for Structural Bioinformatics Protein Data Bank* 2018) existiert. SPARROW wurde auf den Daten der DrugBank (Law *et al.* 2013) ausgeführt, um eine möglichst sichere Datengrundlage zu gewährleisten. Eine beispielhafte Ausgabe von SPARROW für ein privilegiertes Scaffold ist in Abbildung III.3.1b dargestellt. Die vielversprechenden Scaffolds und Proteine, welche durch SPARROW identifiziert wurden, wurden anschließend einer manuellen Verifizierung unterzogen. Eines dieser Scaffolds ist das in der Abbildung III.3.1 gezeigte Chromon-Scaffold, welches als Substruktur in den beiden DrugBank-Molekülen Flavopridol und Genistein vorkommt. Die adressierten Targetproteine sind die CDK2 (Flavopiridol) und der Estrogenrezeptor β (ESR2) (Genistein). Die Proteine sind weder funktionell verwandt, noch weisen sie eine hohe Sequenzähnlichkeit auf. Die Alignments befinden sich im Anhang unter Abschnitt B.1. Somit lieferte SPARROW einen guten Startpunkt zur Überprüfung des APSC, welche im Folgenden näher beschrieben wird.

III.3.2.2. Anwendung des APSC auf das Chromon-Scaffold

Das durch SPARROW gefundene privilegierte Chromon-Scaffold konnte als Substruktur in Genistein (Abbildung III.3.2b) und in Flavopiridol (Abbildung III.3.2a) sowie ebenfalls in weiteren bioaktiven Molekülen, wie Quercetin (Abbildung III.3.2d) und Luteolin (Abbildung III.3.2c) gefunden werden. Es fällt auf, dass es sich bei den genannten Verbindungen um Polyphenole und Naturstoffe bzw. im Falle von Falvopiridol um ein Derivat eines Naturstoffs (Rohitukin) handelt. Diese Verbindungen weisen Eigenschaften auf, die bei der Beurteilung von Assay-Ergebnissen besonderer Aufmerksamkeit bedürfen. Sie können auf verschiedene Weisen mit dem Assaysytem interferieren, zum einen sind Quercetin und Luteolin bekannte Aggregatoren (Irwin *et al.* 2015; *Aggregator Advisor* 2019) und zum anderen können sie durch das konjugierte π-Elektronensystem das optische Read-out von Assays beeinflussen. Zudem sind sie häufig redoxaktiv sowie Metallchelatoren (Jadhav *et al.* 2010) und außerdem sind Substrukturen (z. B. 1,2-Dihydroxybenzen) einiger Polyphenole als *pan-assay interference structures* (PAINS) (Baell 2016) bekannt.



(a) Scaffold Baum des Chromon-Scaffolds



(b) Ausgabe aus SPARROW für das Chromon-Scaffold

Abbildung III.3.1.: Ansicht des Scaffold Baums für das Chromon-Scaffold von Scaffold Hunter (Schäfer *et al.* 2017) (a) und dreidimensionaler Plot als Ausgabe von SPARROW für das Chromon-Scaffold (b). Beide Ausgaben wurden zur besseren Visualisierung leicht verändert.

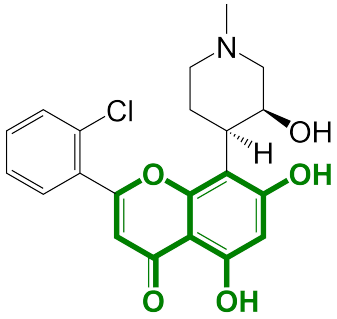
Nichtsdestotrotz liefern Polyphenole einen wertvollen ersten Ausgangspunkt von dem aus durch Optimierung des pharmakokinetischen Profils, der Effizienz und des Selektivitätsprofils Derivate wie Flavopiridol oder RVX-208 (Struktur in Abbildung II.0.5a, Derivat von Resveratrol) entwickelt werden können. Beide Verbindungen befinden sich aktuell in klinischen Studien, wobei Flavopiridol in Phase II und RVX-208 in Phase III getestet wird (*ClinicalTrials* 2019).

Zur Anwendung des APSC auf das Chromon-Scaffold sollten weitere Targetproteine des Scaffolds durch Datenbankenrecherchen z. B. in der PDB identifiziert werden. So bindet Flavopiridol ebenfalls in der Bindetasche von BRD4 und Genistein, Quercetin sowie Luteolin sind Liganden des PPARγ. Eine Auswahl an bioaktiven Molekülstrukturen mit dem privilegierten Chromon-Scaffold und eine Auswahl ihrer Targetproteine ist in der Abbildung III.3.2 dargestellt. Aus den Arbeiten von Koch und Waldmann (2005) und Balamurugan, Dekker und Waldmann (2005) war eine Ähnlichkeit zwischen PPARγ, ESR2 und dem *bile acid receptor* (Gallensäurerezeptor) (NR1H4) bekannt. Aus diesem Grund wurde die Analyse zum APSC auf den NR1H4 ausgeweitet und umfasste nun die Proteine: BRD4, PPARγ, CDK2, ESR2 sowie NR1H4. Ein Alignment und ein phylogenetischer Baum der fünf Proteinsequenzen befindet sich im Anhang unter B.1.

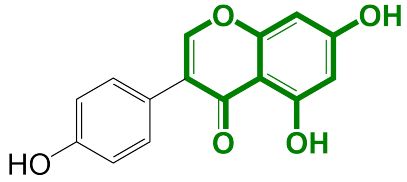
III.3.2.3. Analyse des Chromon-Scaffolds als erweitertes privilegiertes Scaffold

Um den Status des Chromon-Scaffolds als erweitertes privilegiertes Scaffold zu verifizieren, wurde eine genauere Analyse sowohl des Scaffolds als auch der Bindetaschen vorgenommen. Zur Untersuchung der Ähnlichkeiten der Proteine wurden die Kristallstrukturen überlagert (siehe Abbildung III.3.3). Eine weitere Abbildung, welche die Übereinanderlagerung des Scaffolds der BRD4- und ESR2-Komplexstruktur zeigt, befindet sich im Anhang in Abbildung B.2. Es kann eine ähnliche Anordnung von Sekundärstrukturelementen festgestellt werden.

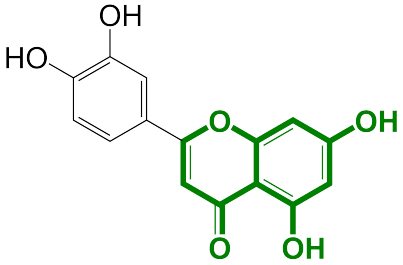
Allerdings stimmt diese Sekundärstruktur-basierte Überlagerung nicht mit einer Überlagerung nach Form und Eigenschaften (Interaktionsfelder) der Bindetaschen überein (vergleiche Abbildungen III.3.3b und III.3.5a). Dies konnte durch eine Analyse der Bindetaschen mittels IsoMIF (Chartier und Najmanovich 2015; Chartier, Adriansen und Najmanovich 2016) gezeigt werden. Einem Subset der sc-PDB (Kellenberger *et al.* 2006; Desaphy *et al.* 2015) wurden die gefundenen Proteine hinzugefügt und BRD4 als Referenzstruktur verwendet. Außerdem wurden zwei verschiedene Ansätze analysiert. Im ersten Ansatz wurden die vollständigen Liganden (kurz Ligand) zur Bindetaschendefinition verwendet, während im zweiten Ansatz nur ein Scaffold (kurz Scaffold) verwendet wurde. Die Ergebnisse der IsoMIF-Suche nach zu BRD4 ähnlichen Bindetaschen sind in einem Histogramm in der Abbildung III.3.4 zusammengefasst. Es ist deutlich erkennbar, dass für die Scaffold-basierte Analyse die Scores aller vier APSC-Proteine oberhalb der zufälligen Bindetaschenähnlichkeiten angesiedelt sind.



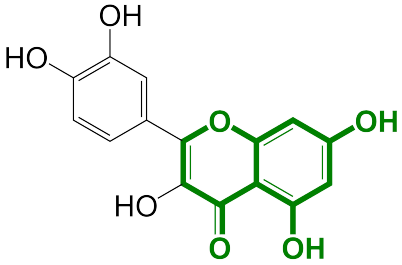
(a) Flavopiridol
Targets: CDK1, CDK2, CDK4-9 (CDK9 PDB-ID: 3BLR), BRD4 (PDB-ID: 4O71), EGFR



(b) Genistein
Targets: ESR1, ESR2 (PDB-IDs: 1X7J, 1QKM), PPARγ

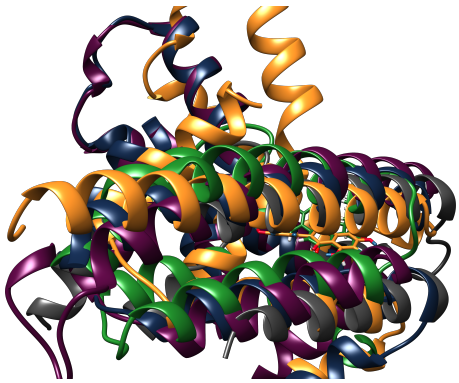


(c) Luteolin
Target: PPARγ (PDB-ID: 3SZ1)

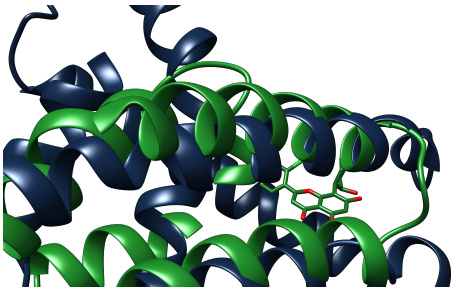


(d) Quercetin
Targets: ESR1, ESR2, PPARγ

Abbildung III.3.2.: Bioaktive Moleküle mit Chromon-Scaffold: Flavopridol (a), Genistein (b), Luteolin (c) und Quercetin (d). In grün ist ein gemeinsames Scaffold (5,7-Dihydroxychromon) hervorgehoben, welches eine Erweiterung des privilegierten Chromon-Scaffolds darstellt. Die Informationen zu einigen Targetproteinen stammen aus der DrugBank (Law *et al.* 2013), der PDB (Berman *et al.* 2000) und Wang *et al.* (2014).



(a) Übereinanderlagerung aller untersuchten APSC Proteine



(b) Übereinanderlagerung von BRD4 (grün) und PPARγ (blau)

Abbildung III.3.3.: Strukturen von BRD4 (grün, PDB-ID: 4O71), PPARγ (blau, PDB-ID: 1NYX), CDK2 (grau, PDB-ID: 1GZ8), ESR2 (orange, PDB-ID: 1QKM) sowie NR1H4 (weinrot, PDB-ID: 3BEJ).

Ähnliches gilt für den Ligand-basierten Ansatz, jedoch fällt der Abstand zur zufälligen Verteilung der Ähnlichkeiten etwas geringer aus. Für NR1H4 liegt die Bindetaschenähnlichkeit zu BRD4 basierend auf vollständigen Liganden sogar innerhalb der zufälligen Verteilung. Anhand der großen Streubreite von Bindetaschenähnlichkeiten von BRD4 zu sc-PDB-Strukturen, welche Quercetin (Abbildung III.3.2d) als Liganden enthalten und den im Vergleich zur BRD4-PPARγ-Bindetaschenähnlichkeit geringen Ähnlichkeiten, kann geschlussfolgert werden, dass die beobachtete Bindetaschenähnlichkeit zwischen BRD4 und PPARγ nicht ausschließlich auf eine Ähnlichkeit der Liganden zurückzuführen ist. Die Streubreite bleibt ebenfalls erhalten, wenn nur das Scaffold des Liganden der BRD4-Komplexstruktur betrachtet wird. Zur weiteren Verifizierung wurden dem sc-PDB-Datensatz ebenfalls Luteolin-Komplexstrukturen sowie eine Flavopiridol-Komplexstruktur hinzugefügt, da Luteolin (Abbildung III.3.2c) der Ligand der verwendeten PPARγ-Komplexstruktur ist und Flavopiridol (Abbildung III.3.2a) der Ligand der BRD4-Komplexstruktur ist, welche als Suchstruktur verwendet wurde. Alle drei Liganden (Quercetin, Luteolin und Flavopiridol) enthalten zudem das Chromon-Scaffold. Die Flavopiridol-Komplexstruktur (PDB-ID: 3EBP) erhält für beide Ansätze einen deutlich geringeren Score als die APSC-Proteine mit Ausnahme von NR1H4 für den Ansatz mit vollständigen Liganden. Die Ergebnisse der Luteolin-Komplexstrukturen zeigen, ähnlich wie die Quercetin-Ergebnisse, eine große Streubreite (siehe Abbildung III.3.4). Interessant ist, dass Luteolin-Komplexstrukturen von Kinasen (PDB-IDs: 5AUU und 4DGN) oder Bromodomänen (PDB-ID: 5II2) höhere Scores als die anderen Luteolin-Komplexstrukturen erzielen. Insgesamt erzielten die Kinase-Strukturen (PDB-IDs: 5AUU, 4DGN, 1CX4) und die Bromodomän-Strukturen (PDB-IDs: 5II2 und 3U5L) hohe Scores.

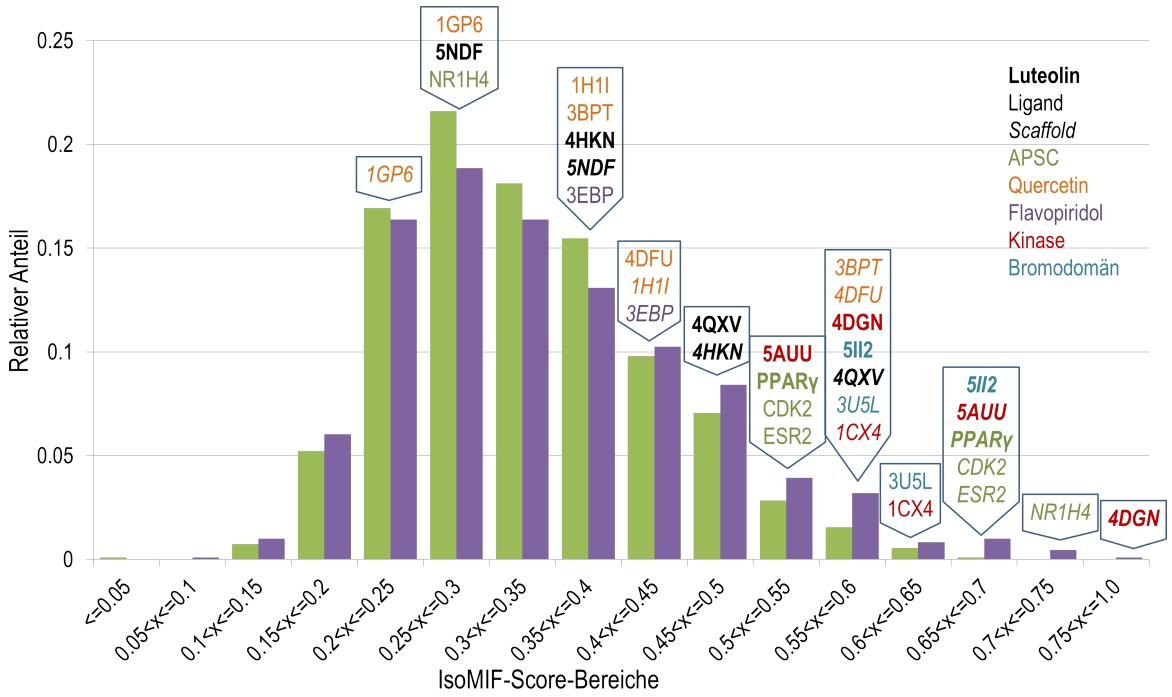


Abbildung III.3.4.: Ergebnis des Bindetaschenvergleichs mit IsoMIF (Chartier, Adriansen und Najmanovich 2016) von BRD4 (PDB-ID: 4O71) gegen ein Subset der sc-PDB inklusive der vier anderen Strukturen des APSC, deren Ergebnisse grün hervorgehoben sind. Die Bindetasche wurde entweder mithilfe des vollständigen Liganden (Liganden, grüne Balken) oder nur mithilfe des Scaffolds (Scaffold, lila Balken, kursive PDB-IDs) definiert. Zudem wurden die Ergebnisse für Proteinstrukturen, welche als Liganden Quercetin (orange), Flavopiridol (lila) oder Luteolin (fett) enthalten sowie für Strukturen von Kinasen (rot) oder Bromodomänen (türkis) durch Kennzeichnung mit der PDB-ID hervorgehoben.

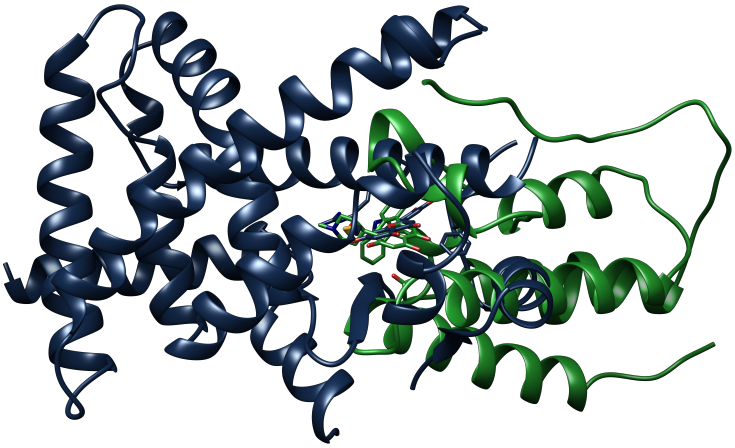
Die Ergebnisse sind im Einklang mit vorherigen Analysen, welche Ähnlichkeiten von BET-Proteinen und Kinasen basierend auf ähnlichen Liganden aufzeigen (Carlino und Rastelli 2016; Ember *et al.* 2014). Zusammenfassend lässt sich schlussfolgern, dass der IsoMIF-Score stärker von der Proteinähnlichkeit als von der Ligandähnlichkeit abhängt und die Ähnlichkeit zwischen BRD4 und den APSC-Proteinen mit Ausnahme von NR1H4 bedeutend ist.

Anschließend an die Score-Analyse wurden die durch IsoMIF generierten Alignments untersucht. In beiden Ansätzen (vollständige Liganden und Scaffold) konnte das Scaffold nur für BRD4 und PPARγ gut durch IsoMIF überlagert werden (siehe Abbildungen III.3.5 und B.3 im Anhang). Dies lässt eine besondere Verbindung von PPARγ und BRD4 vermuten und legt eine gemeinsame Erkennung von ähnlichen Liganden nahe.

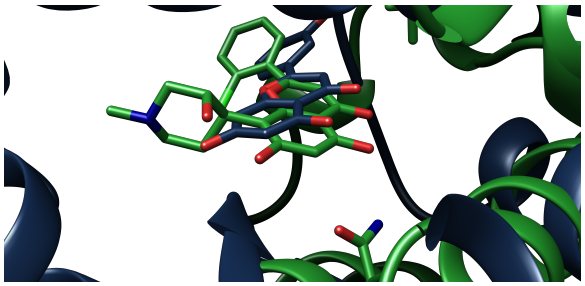
Zudem bildet die ähnliche Anordnung der Liganden in dem IsoMIF-basierten Alignment eine rationale Grundlage für die Akivität von Liganden gegenüber beiden Proteinen und erlaubt somit ein zielgerichtetes Design von dualen Modulatoren oder ein Wiederverwenden (*repositioning*) von Liganden des einen Proteins z. B. PPARγ zur Beeinflussung des anderen Proteins z. B. BRD4. Außerdem kann das Alignment das rationalen Design von selektiven Liganden unterstützen. Leichte Abweichungen der in den ESR2- und NR1H4-Komplexstrukturen enthaltenen Scaffolds vom Scaffold der BRD4-Komplexstruktur könnten eine Erklärung für die unzureichende Übereinanderlagerung dieser liefern. Beim ESR2-Scaffold handelt es sich statt um ein 2-Phenyl-4H-chromen um ein 3-Phenyl-4H-chromen. Das NR1H4-Scaffold besitzt anstelle des Chromon-Ringsystems ein Naphthalin-Ringsystem. In den Alignments beider Ansätze sind die Scaffolds der BRD4-, PPARγ-, CDK2- und ESR2-Kristallstrukturen (Abbildungen III.3.5b, III.3.5c und III.3.5d sowie B.3b, III.3.5c und III.3.5d im Anhang) ähnlich ausgerichtet, d. h. in ungefähr einer Ebene, wobei die ähnliche Ausrichtung deutlicher im Scaffold-basierten Alignment erkennbar ist. Für die Scaffolds der BRD4- und ESR2-Komplexstrukturen scheint ein besseres Scaffold-basiertes Alignment denkbar, wenn es sich beim ESR2-Scaffold ebenfalls um ein 2-Phenyl-4H-chromen handeln würde. Generell ist anzumerken, dass nur ein Ausschnitt aus möglichen Alignments untersucht wurde, einerseits begrenzt durch die verfügbaren Kokristallstrukturen und andererseits durch die Konzentration auf die APSC-Proteine. Bei einer Erweiterung auf ähnliche Proteine in Komplex mit ähnlichen Liganden (siehe unten) wären weitere Alignments möglich. Demzufolge kann aufgrund dieser Analyse eine besondere Verbindung von BRD4 und PPARγ angenommen werden, jedoch kann diese nicht für die andere Kombinationen ausgeschlossen werden.

Zusammenfassend wurde eine Ähnlichkeit der Bindetaschen von BRD4 und PPARγ durch das APSC entdeckt, welche über eine zufällige Bindetaschenähnlichkeit hinausgeht und für das rationale Wirkstoffdesign bezüglich BRD4 und PPARγ von Bedeutung ist. Dennoch können selektive Liganden entwickelt werden. Ein Beispiel für zwei selektive Liganden ist im Anhang in der Abbildung B.4 gezeigt. Die selektiven Liganden zeigen nur Wasserstoffbrücken-vermittelte Interaktionen zu dem Protein, in dem sie kokristallisiert wurden.

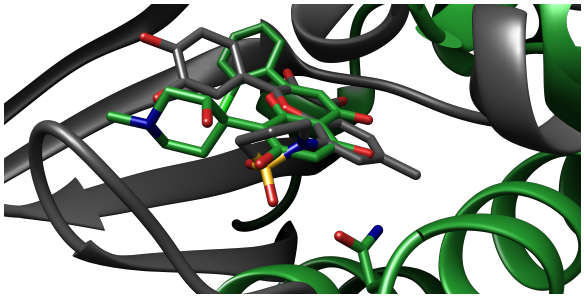
Das Chromon-Scaffold und insbesondere seine Erweiterung (siehe grün hervorgehobene Substruktur in Abbildung III.3.2) bietet viele Interaktionsmöglichkeiten. Ein besonderes Merkmal ist dabei die Anordnung der Substituenten, da sie verschiedene Bindemodi des Scaffolds trotz ähnlicher/gleicher Interaktionen ermöglicht. Einige Bespiele sind in den Abbildungen B.5 für BRDs und B.6 für CDK2 und CDK9 im Anhang abgebildet.



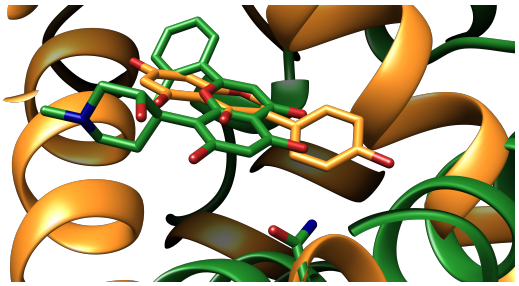
(a) Übersicht BRD4 (grün) und PPARγ (blau)



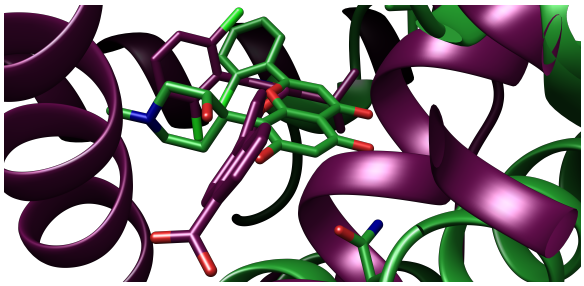
(b) BRD4 (grün) und PPARγ (blau)



(c) BRD4 (grün) und CDK2 (grau)



(d) BRD4 (grün) und ESR2 (orange)



(e) BRD4 (grün) und NR1H4 (weinrot)

Abbildung III.3.5.: Alignment von BRD4 (grün, PDB-ID: 4O71) mit PPARγ (blau, PDB-ID: 3SZ1) als Übersicht (a) und Nahaufnahme (b), mit CDK2 (grau, PDB-ID: 2DUV) (c), mit ESR2 (orange, PDB-ID: 1QKM) (d) sowie mit NR1H4 (weinrot, PDB-ID: 3BEJ) (e) mittels IsoMIF basierend auf vollständigen Liganden.

III.3.2.4. Ergebnisse des initialen Screens

Nach den positiven Ergebnissen der vorhergehenden Analysen, sollte das APSC nun biochemisch verifiziert werden. Dazu wurden die Liganden der Proteine PPARγ, CDK2, ESR2 sowie NR1H4 aus der ChEMBL (Bento *et al.* 2014) mittels eines Pipeline Pilot Workflows extrahiert. Anschließend wurde in der internen Substanzbibliothek des *Compound Management and Screening Center* (COMAS) virtuell nach ähnlichen Molekülen zu den gefundenen Liganden aus der ChEMBL gesucht. Es wurden insgesamt 1158 ähnliche Moleküle aus der COMAS-Substanzbibliothek erhalten und in einem initialen Screen per HTRF-Assay (siehe Abschnitt IV.1.5.1) auf ihre biologische Aktivität gegenüber BRD4 bei einer Konzentration von 10 μM getestet. Zur Kontrolle und zum Vergleich wurden parallel zum und unabhängig vom APSC Verbindungen zur Testung auf BRD4-Inhibition mittels eines Docking-Ansatzes ausgewählt. Die Ergebnisse sind in Tabelle III.3.1 zusammengefasst. Dabei wurde ein Hit definiert als ein Analyt, der die verbleibende Bindungsaktiviät (kurz Restaktivität) auf 70 % ($Hits_{70}$) oder 50 % ($Hits_{50}$) senkt.

	# $Hits_{50}$	# $Hits_{70}$	# bestätigte Hits	% $Hits_{50}$	% $Hits_{70}$	$\sum$ getestet
CDK2	1	3	1	1,22	3,67	82
ESR2	5	17	4	0,66	2,24	780
PPARγ	25	40	14	8,96	14,34	279
NR1H4	0	0	0	0	0	39
$\sum$ APSC	31	60	19	2,68	5,18	1158
Docking	41	87	23	2,58	5,46	1592

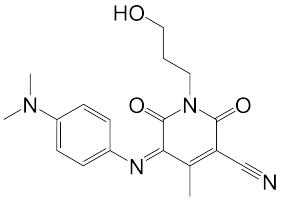
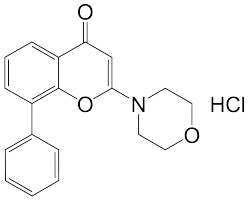
Tabelle III.3.1.: Ergebnisse des initialen BRD4-Screens bei einer Analytenkonzentration von 10 μM. Angegeben sind die Anzahl und der Prozentsatz an Hits aufgeschlüsselt nach Ursprungsprotein bzw. Methode. Dabei wurde ein Hit definiert als ein Analyt, der die verbleibende Bindungsaktiviät (kurz Restaktivität) auf 70 % ($Hits_{70}$) oder 50 % ($Hits_{50}$) senkt.

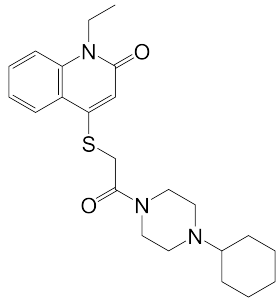
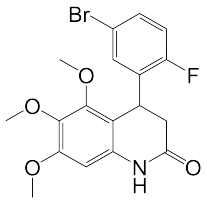
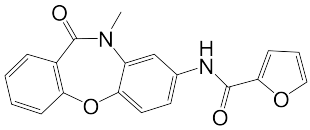
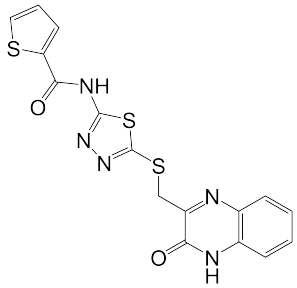
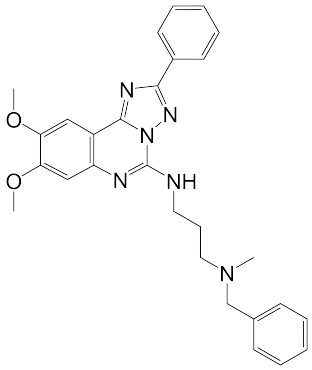
Es ist bekannt, dass einige Kinase-Inhibitoren ebenfalls BRD4 inhibieren (Ember *et al.* 2014; Carlino und Rastelli 2016). Dieser bereits publizierte Zusammenhang konnte ebenfalls mithilfe des neuen APSC gezeigt werden. Somit erfolgte eine erste Bestätigung des Konzeptes. Wirklich bemerkenswert ist darüberhinaus der auffallend hohe Anteil (14,34 %) an BRD4-$Hits_{70}$, welche ähnlich zu PPARγ-Liganden sind. Aus diesem Grund wurde im Folgenden auf die beiden Proteine BRD4 und PPARγ fokussiert. Eine Übersicht der physiologischen Funktionen sowie der pathophysiologischen Bedeutung der Proteine befindet sich unter dem Abschnitt II.2. Die durchschnittliche Hit_{70}-Rate des APSC betrug 5,18 % und ist somit vergleichbar mit der Hit_{70}-Rate von 5,46 % des Docking-basierten Ansatzes. Allerdings unterscheiden sich die in den beiden Ansätzen jeweils gefundenen Hits.

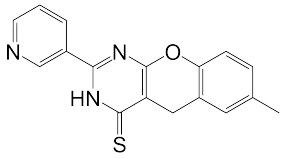
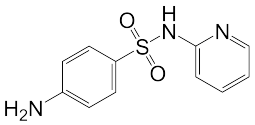
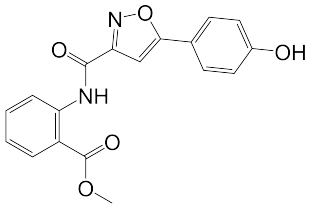
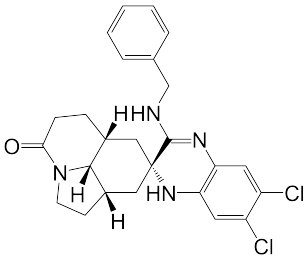
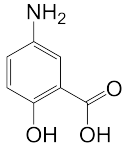
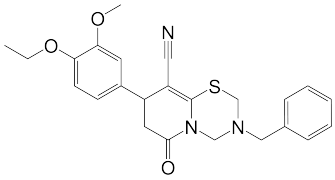
Eine Ähnlichkeitsanalyse mithilfe des ECFP4 und des TC der Hits aus dem APSC und dem Docking-Ansatz sowie ein Vergleich dieser Hits zu bekannten BRD4- und PPARγ-Liganden befindet sich inklusive einer kurzen Diskussion im Anhang in der Abbildung B.7 und unter dem Abschnitt B.3. Die zweithöchste Hit_{70}-Rate, nach der für Analyten mit Ursprungsprotein PPARγ, konnte mit 3,67 % für Analyten mit dem Ursprungsprotein CDK2 erzielt werden. Danach folgen mit einer Hit_{70}-Rate von 2,24 % die Analyten mit dem Ursprungsprotein ESR2. Es konnte kein Hit unter den Analyten mit Ursprungsprotein NR1H4 identifiziert werden. Dies könnte an der geringen Anzahl an getesteten Verbindungen (39) liegen. Bei einer angenommenen Hit_{70}-Rate von 2 % für Analyten mit Ursprungsprotein NR1H4 hätten theoretisch 0,78 Hits gefunden werden sollen. Zudem konnte IsoMIF keine Bindetaschenähnlichkeit zwischen NR1H4 und BRD4 basierend auf vollständigen Liganden finden, welche sich von der zufälligen Verteilung von Bindetaschenähnlichkeiten zwischen BRD4 und einem Subset von Proteinen aus der sc-PDB unterscheidet (siehe Abschnitt III.3.2.3 und Abbildung III.3.4). Die von IsoMIF generierten Alignments von BRD4 und NR1H4 weisen ebenfalls eine unzureichende Übereinanderlagerung des Scaffolds auf. Im Folgenden soll eine biochemische Charakterisierung ausgewählter initialer Hits erfolgen.

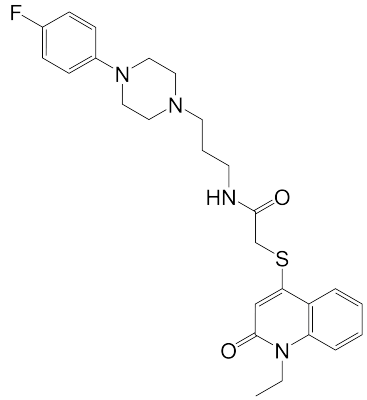
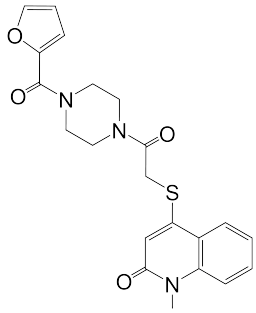
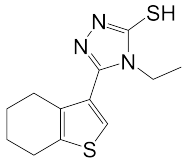
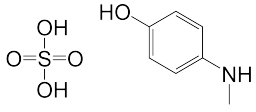
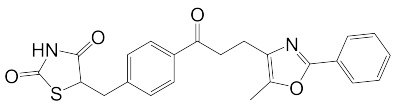
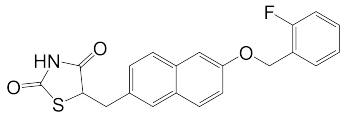
III.3.2.5. Biochemische Charakterisierung initialer Hits

Zur biochemischen Charakterisierung initialer Hits wurde der IC_{50}-Wert mithilfe des HTRF-Assays bestimmt. Die Ergebnisse sind in der Tabelle III.3.2 zusammengefasst. Im Folgenden werden die Symbole Ø für den Durchschnitt und σ für die Standardabweichung verwendet.

ID	Struktur	Ø IC_{50} $\pm$ σ in μM	Ø K_d $\pm$ σ in μM	Ø ΔTm in °C	PDB-ID	Quelle	Trivialname oder Bemerkung
1		1,64 $\pm$ 0,62	1,47 $\pm$ 0,06	0,47		APSC PPARγ	PAINS Strukturen[†]
2		5,1 $\pm$ 0,59	4,08 $\pm$ 0,33	1,86***	6G0D	APSC CDK2	

ID	Struktur	Ø IC_{50} $\pm$ σ in μM	Ø K_d	Ø ΔTm in ℃	PDB-ID	Quelle	Trivialname oder Bemerkung
3		5,62 ± 0,65	4,83 ± 0,35	1,63***	6G0E	APSC PPARγ	
4		0,58 ± 0,08	1,64 ± 0,03	1,55***	6G0F	Docking	
5		2,20 ± 0,17	2,8 ± 0,6	1,30***		Docking	
6		5,96 ± 2,71	n. a.	-4,55***		Docking	falsch positiv, destabilisiert BRD4
7		3,11 ± 0,52	2,70 ± 0,23	0,41		Docking	

ID	Struktur	Ø IC_{50} $\pm$ σ in μM	Ø K_d	Ø ΔTm in ℃	PDB-ID	Quelle	Trivialname oder Bemerkung
8		7,03 ± 2,73	16,88 ± 2,42	0,63		Docking	
9		> 10	7,32 ± 0,35	1,01***		Sulfa	falsch negativ, Sulfasalazin-Substruktur
10		5,08 ± 0,40	> 100	1,00**		APSC PPARγ	falsch positiv
11		8,84 ± 0,96	10,27 ± 1,10	1,03**		Docking	
12		> 10	> 100	0,78*		Sulfa	Mesalazin
13		> 10	33,65 ± 1,87	0,90**		APSC PPARγ	

ID	Struktur	Ø IC_{50} $\pm$ σ in μM	Ø K_d	Ø ΔTm in ℃	PDB-ID	Quelle	Trivialname oder Bemerkung
14		8,01 ± 0,47	7,77 ± 0,13	1,01***		APSC PPARγ	
15		> 10	87,43 ± 0,97	1,00***		APSC PPARγ	
16		> 10	104,70 ± 3,71	1,07***		APSC PPARγ	
17		2,410 ± 0,37	n. a.	n. a.		APSC PPARγ	falsch positiv, destabilisiert BRD4
18		> 10	46,58 ± 0,82	-0,28		PPARγ-Agonist	Darglitazon
19		> 10	> 100	0,28		PPARγ-Modulator	Netoglitazon

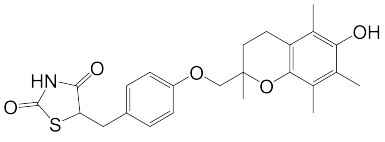
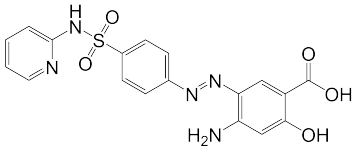
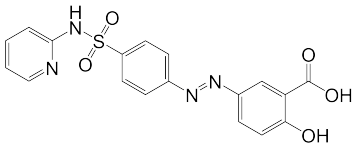
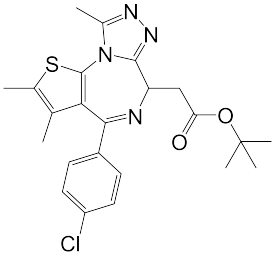
ID	Struktur	Ø IC_{50} ± σ in μM	Ø K_d	Ø ΔTm in ℃	PDB-ID	Quelle	Trivialname oder Bemerkung
20		> 10	> 100	n. g.		PPARγ-Agonist	Troglitazon
21		> 10	> 100	n. g.	6G0H	Sulfa	SSLH01
22		> 10	> 100	n. g.	6G0G	Sulfa	Sulfasalazin
23		0,06 ± 0,01	n. g.	5,34***	3MXF‡	Kontrolle	± JQ1

Tabelle III.3.2.: Ergebnisse ausgewählter initialer Hits. Alle Werte wurden als Mittelwert aus drei unabhähnigen Messungen bestimmt. IC_{50}-Werte wurden mittels HTRF-Assay, K_d-Werte per MST und ΔTm-Werte durch den TSA ermittelt. Eine einfaktorielle Varianzanalyse mit Dunnets-Methode für Mehrfachvergleiche wurde durchgeführt, um die Mittelwerte der Schmelztemperaturen bei Analytenzusatz und der Protein-Kontrolle (ohne Zusatz eines Analyten) nach Abzug der Analyten-Kontrolle und der ohne-Protein-Kontrolle zu vergleichen. * $p < 0{,}33$, ** $p < 0{,}002$, *** $p < 0{,}001$, n. a.: nicht auswertbar, n. g.: nicht gemessen, Tm: Schmelztemperatur, †: ist jedoch kein Hit in anderen ähnlichen Assays, ‡: Referenz: Filippakopoulos *et al.* (2010), Sulfa: zusätzliche Verbindungen zur Untersuchung von Sulfasalazin und Derivaten.

Im Folgenden sollen zwei Substanzserien separat detaillierter betrachtet werden.

III.3.2.6. Biochemische Charakterisierung einer Substanzserie zur Inhibition von EGFR-Varianten

Im Primärscreen wurde ein Hit aus einer Substanzserie zur Inhibition von *epidermal growth factor receptor* (Epidermaler Wachstumsfaktorrezeptor) (EGFR) Varianten, welche in der Arbeitsgruppe von Prof. Rauh an der TU Dortmund entwickelt wurde, identifiziert (RL129). EGFR ist eine Rezeptortyrosinkinase, welche für die Regulation von u. a. Zellwachstum, Proliferation und Differenzierung eine wichtige Rolle spielt (Wieduwilt und Moasser 2008). Eine erhöhte, ligandunabhänige Kinaseaktivität durch aktivierende Mutationen gilt als ein onkogener Treiber (*oncogenic driver*) des nicht-kleinzelligen Lungenkarzinoms. Damit stellt EGFR einen wichtigen Angriffspunkt für die zielgerichtete und personalisierte Therapie des nicht-kleinzelligen Lungenkarzinoms dar. In der Arbeitsgruppe von Prof. Rauh werden EGFR und seine Varianten intensiv erforscht (Engel *et al.* 2015; Engel *et al.* 2016; Engel *et al.* 2017; Smith *et al.* 2017; Günther *et al.* 2017; Tomassi *et al.* 2017). Die Substanzserie beruht auf Vorarbeiten von Getlik (2011) und Engel (2018). Interessanterweise ist ein dualer EGFR-BRD4-Inhibitor (Verbindung 2870, siehe Abbildung III.3.6), welcher ebenfalls ein 4-Aminochinazolin-Scaffold besitzt, bekannt (Allen *et al.* 2015). Ziel dieser dualen Inhibitoren ist, die Effizienz zu steigern und eine Resistenzentwicklung zu verhindern. Neben dualen EGFR-BRD4-Inhibitoren zeigen u. a. duale EGFR-PI3K-Inhibitoren (PIK: Phosphoinositid-3-Kinase) ebenfalls ein entsprechendes Potential (Sos *et al.* 2010).

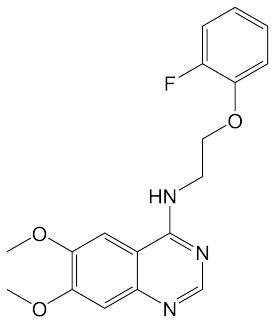


Abbildung III.3.6.: Bekannter dualer EGFR-BRD4-Inhibitor (Verbindung 2870) mit 4-Aminochinazolin-Scaffold (Allen *et al.* 2015).

Zur Identifizierung weiterer dualer EGFR-BRD4-Inhibitoren wurden weitere 25 Verbindungen aus der EGFR-Inhibitor-Serie auf ihre inhibitorische Wirkung bezüglich BRD4 mittels HTRF-Assay untersucht. Es konnten drei Hits identifiziert werden (siehe Tabelle III.3.3). Davon konnte für zwei Verbindungen ein IC_{50}-Wert unter der Verwendung einer Konzentrationsreihe (von 3 nM bis 10 μM) ermittelt werden. Alle getesteten Verbindungen sowie deren Ergebnisse sind in Tabelle B.5 im Anhang zusammengefasst.

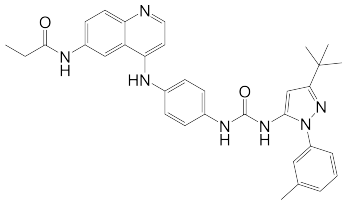
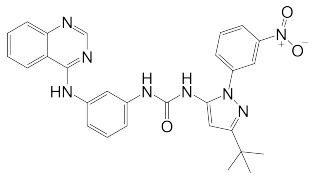
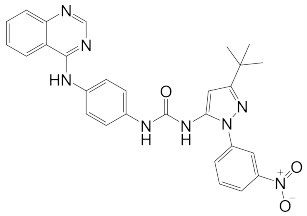
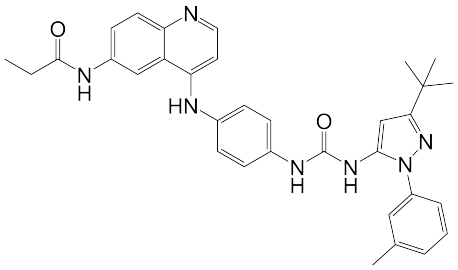
ID	Struktur	Ø $IC_{50} \pm \sigma$ in μM
RL078		5,36 ± 0,18
RL127		> 10
RL129		7,92 ± 0,07

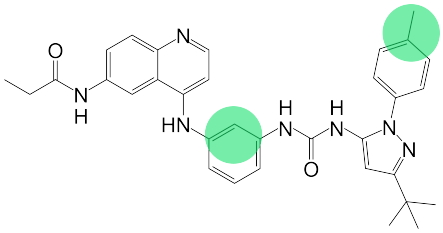
Tabelle III.3.3.: BRD4-Hits aus der EGFR-Inhibitor-Serie und ermittelte IC_{50}-Werte bei einer Konzentration bis 10 μM.

Anschließend wurden die Strukturen von zu den beiden Inhibitoren RL078 und RL129 ähnlichen Verbindungen, welche jedoch keine Aktivität im HTRF-Assay zeigten, analysiert. Damit sollten erste Einblicke in die Struktur-Wirkungsbeziehung erlangt werden. Die Strukturen sind in den Abbildungen III.3.7 (für RL078) und III.3.8 (für RL129) dargestellt.

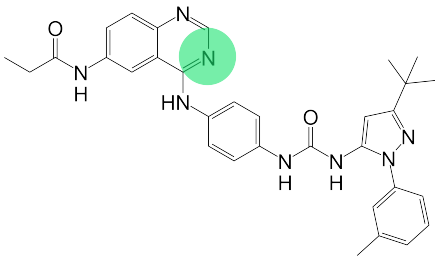
Die meta-Stellung am mittleren Benzen-Ring (siehe RL127 Tabelle III.3.3 und RL077 Abbildung III.3.7b) scheint gegenüber der para-Stellung (siehe RL129 Tabelle III.3.3 und RL078 Abbildung III.3.7a) benachteiligt zu sein. Am endständigen Benzen-Ring weisen hingegen alle aktiven Verbindungen eine meta-Stellung der Substituenten auf. Der Austausch des Chinolin-Grundgerüstes (Abbildung III.3.7a) gegen das Chinazolin-Grundgerüst (Abbildung III.3.7c) führt zu einer inaktiven Verbindung. Da das Chinazolin-Grundgerüst jedoch in der aktiven Verbindung RL129 (Abbildung III.3.8a) enthalten ist, ist zu vermuten, dass eine Kombination von Faktoren zum Aktivitätsverlust führt. Eine Erklärung wäre eine Kombination aus einer erhöhten Raumbeanspruchung durch den Propanamid-Substituenten (siehe auch Abbildung III.3.8) und einem veränderten Bindemodus, hervorgerufen durch das Chinazolin-Grundgerüst.



(a) RL078
IC_{50} = 5,36 μM

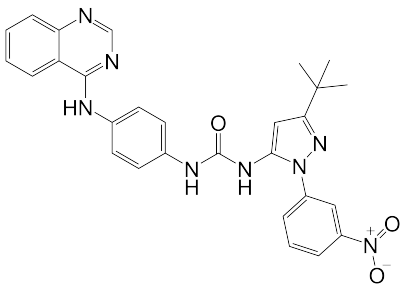


(b) RL077
inaktiv

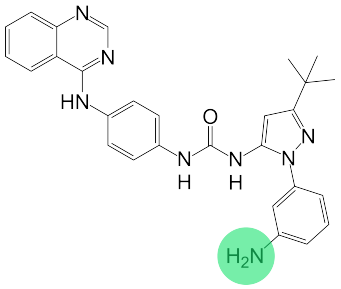


(c) RL046
inaktiv

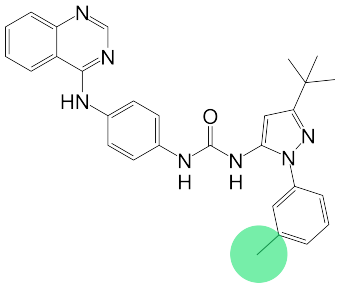
Abbildung III.3.7.: Strukturen von RL078 (a), RL077 (b) und RL046 (c).



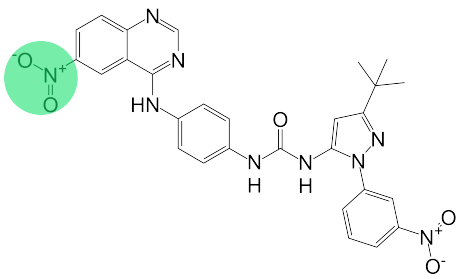
(a) RL129
IC_{50} = 7,92 μM



(b) RL130
inaktiv



(c) RL070
inaktiv



(d) RL100
inaktiv

Abbildung III.3.8.: Strukturen von RL129 (a), RL130 (b), RL070 (c) und RL100 (d).

Bei der Analyse der zu der aktiven Verbindung RL129 ähnlichen Verbindungen RL130 und RL070 (siehe Abbildung III.3.8) fällt auf, dass scheinbar eine Veränderung des Substituenten (in RL129 eine Nitrogruppe) in meta-Position der an den Pyrazol-Ring gebundenen Phenylgruppe nicht toleriert wird. Ebenso führt die Einführung einer Nitrogruppe am Chinazolin-Grundgerüst zu einer inaktiven Verbindung (RL100, Abbildung III.3.8d). Der Aktivitätsverlust könnte für RL100 ebenfalls auf einer erhöhten Raumbeanspruchung beruhen. Zusammenfassend kann festgestellt werden, dass in der untersuchten EGFR-Inhibitor-Serie erstaunlicherweise bereits kleine Veränderungen der Struktur zu einer drastischen Änderung der Potenz der Verbindung bezüglich BRD4 führen (*activity cliff*).

Da für die getesteten Verbindungen der EGFR-Inhibitor-Serie zelluläre Daten vorlagen (Brägelmann *et al.* 2017), sollte anschließend an die biochemische Charakterisierung eine Analyse der Effekte auf Krebszelllinien durchgeführt werden.

Analyse der zellulären Effekte der BRD4-Inhibitoren aus der EGFR-Inhibitor-Serie

In einem zellulären Screen wurde der Effekt der Substanzen der EGFR-Inhibitor-Serie auf 92 verschiedene Krebszelllinien bestimmt. Der Schwerpunkt lag dabei auf Lungenkrebszelllinien und besonders auf Zelllinien der speziellen Lungenkrebsform des nichtkleinzelligen Lungenkarzinoms, da in diesen Krebsarten besonders häufig Mutationen im *EGFR*-Gen und erhöhte EGFR-Expressionslevel auftreten. Mit der Analyse der zellulären Effekte soll überprüft werden, ob sich EGFR-Inhibitoren in ihrem Wirkprofil von dualen EGFR-BRD4-Inhibitoren unterscheiden. Zudem sollen mögliche Biomarker-Kanditaten identifiziert werden, um das Ansprechen auf eine duale Inhibition vorhersagen zu können. Es wurden die Zelllinien ermittelt, deren mittlere Proliferation inklusive Standardabweichung durch die beiden BRD4-Inhibitoren RL078 und RL129 stärker beeinträchtigt wurde als im Durchschnitt und nach Abzug der Standardabweichung aller anderen getesteten auf BRD4 inaktiven Verbindungen. Die ermittelten gegenüber EGFR-BRD4-Inhibition sensitiven Zelllinien sind: HCC461, H647, HCC1195, H2347, H1975, H1563, H1993, H3122, H3122, H2228 und HeLa (siehe Tabelle III.3.4).

Die Zelllinien HCC1195 und H2347 weisen eine *NRAS*-Mutation auf. Aus der Literatur ist bekannt, dass BRD4-Inhibitoren das Tumorwachstum von Tumoren mit *NRAS*-Mutation senken (Echevarría-Vargas *et al.* 2018). Dahingegen sind *KRAS*-Mutationen Auslöser von BETi-Resistenzen (Wyce *et al.* 2018). Allerdings liegt bei den beiden aufgeführten Zelllinien mit mutiertem *KRAS*-Gen (HCC461 und H647) keine parallele Misssense-Mutation des *STK11*-Gens vor. Ein Kombination aus einer *KRAS*- und einer *STK11*-Mutation wird mit einer BETi-Resistenz assoziiert, während Zelllinien mit mutiertem *KRAS*-Gen und Wildtyp *STK11*-Gen eine Sensitivität gegenüber BETi zeigen können (Shimamura *et al.* 2013). Eine ausführliche Diskussion zu BETi-Resistenzen befindet sich unter Abschnitt III.3.2.10. Für die Zelllinien H3122 und H2228 ist bekannt, dass sie sensitiv auf BET-Inhibition reagieren (Riveiro *et al.* 2016).

Zelllinie	Ø Proliferation EGFR-Inhibitoren	σ Proliferation EGFR-Inhibitoren	Ø Proliferation - σ Proliferation EGFR-Inhibitoren ($G_{EGFR,min}$)	Ø Proliferation duale Inhibitoren	σ Proliferation duale Inhibitoren	Ø Proliferation + σ Proliferation duale Inhibitoren ($G_{dual,max}$)	$G_{EGFR,min}$ - $G_{dual,max}$
HCC461	93,97	8,35	85,62	78,93	5,31	84,24	1,38
H647*	84,44	16,19	68,25	47,72	6,77	54,49	13,76
HCC1195*	111,62	30,94	80,68	65,79	3,52	69,31	11,37
H2347	90,46	15,41	75,05	70,33	3,33	73,66	1,39
H1975*	99,71	6,63	93,08	86,55	0,38	86,93	6,15
H1563*	78,53	14,51	64,02	54,73	4,53	59,26	4,76
H1993	97,74	10,13	87,61	56,44	29,53	85,97	1,64
H3122	104,74	25,51	79,23	55,97	21,16	77,13	2,10
H3122*	113,69	24,09	89,60	62,69	1,49	64,18	25,42
H2228*	91,93	16,25	75,68	59,25	5,33	64,58	11,10
HeLa	96,32	10,86	85,46	40.00	34,13	74,13	11,33

Tabelle III.3.4.: Zelluläre Charakterisierung von BRD4-Inhibitoren aus der EGFR-Inhibitor-Serie. *: Unterschied zwischen sensitiven und weniger sensitiven Zelllinien ist ebenfalls bei Berechnung der mittleren Proliferation über alle Inhibitoren, d. h. auch RL078 und RL129, messbar. Damit kennzeichnet das * die besonders sensitiven Zelllinien. Die Proliferation ist in Prozent angegeben.

Interessanterweise erfolgt eine unterdurchschnittliche Proliferationshemmung durch RL078 und RL129 gegenüber den BETi resistenten Zelllinien H441 (Proliferation: 93,6 % und 84,9 %), H460 (77,7 % und 80,5 %) und H661 (66,6 %/51,4 % und 83,3 %/84,3 %). Die durchschnittliche Proliferation beträgt 51,2 % für RL078 und 78,7 % für RL129. Allerdings unterscheiden sich RL078 und RL129 in dem Ausmaß in dem ihre proliferationshemmende Wirkung beeinflusst ist für andere BETi resistente Zelllinien. So ist die proliferationshemmende Wirkung von RL078 auf die BETi resistente Zelllinie H520 weiterhin gegeben (5,2 %), während sie für RL129 durchschnittlich (76,5 %) ist. Andererseits ist die BETi resistente Zelllinie A549 gegenüber RL129 sensitiv (61,4 %) und unempfindlich gegenüber RL078 (51,4 %).

Daraus lässt sich schließen, dass der beobachtete Effekt auf die Zellproliferation sowohl auf die EGFR-Inhibition als auch auf die BRD4-Inhibition zurückzuführen ist.

Ein Vergleich der Mutationen in bezüglich der EGFR-BRD4-Inhibition sensitiven und weniger sensitiven, im Folgenden insensitiv genannten, Zelllinien wurde im Rahmen dieser Arbeit mithilfe der *Cancer Cell Line Encyclopedia* (2018) (Cancer Cell Line Encyclopedia Consortium und Genomics of Drug Sensitivity in Cancer Consortium 2015) durchgeführt. Dazu wurden nur die besonders sensitiven Zelllinien (mit * markierte Zelllinien in Tabelle III.3.4) verwendet. Die Mutationen, welche mindestens in der Hälfte der für EGFR-BRD4-Inhibition besonders sensitiven Zelllinien vorkommen, sind zusammen mit ihrem Rang unter Sortierung nach Häufigkeit sowohl für besonders sensitive als auch für insensitive Zelllinien in Tabelle B.6 im Anhang angegeben. In allen sechs besonders sensitiven Zelllinien liegt eine Mutation des *MT-CYB*-Gens vor. In den insensitiven Zelllinien trägt nur ca. die Hälfte eine Mutation des *MT-CYB*-Gens. Die Mutationen mit den drei größten Unterschieden zwischen dem Rang der Häufigkeit in besonders sensitiven versus insensitiven Zelllinien sind in den Genen: *FAM84A* (4 vs. 35), *BCAS3* (4 vs. 31) und *KIAA1462* (4 vs. 30) zu finden. Dabei wurden nur Mutationen betrachtet, die mindestens in 50 % der besonders sensitiven Zelllinien vorlagen. Interessant ist ebenfalls, dass eine Mutation des *RYR2*-Gens in ca. 50 % der insensitiven Zelllinien vorliegt, wohingegen diese Mutation in keiner der sechs besonders sensitiven Zelllinien zu finden ist. Somit gibt der Status des *RYR2*-Gens erste Auskünfte über eine mögliche Resistenz gegenüber den EGFR-BRD4-Inhibitoren und sollte als potentieller Kandidat für einen Biomarker weiter untersucht werden. Für eine genauere Analyse sollten in Zukunft ebenfalls die Expressionslevel und weitere Veränderungen wie Kopienzahlvariationen (*copy number variation*) relevanter Proteine zwischen sensitiven und insensitven Zelllinien verglichen werden.

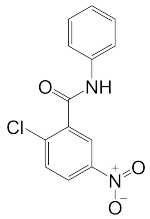
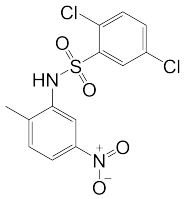
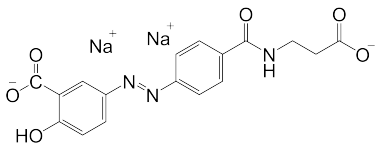
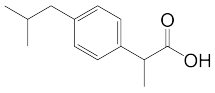
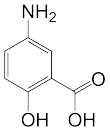
Anzumerken ist, dass die biochemischen Testungen nur für die Bromodomäne 1 (BD1) von BRD4 (kurz: BRD4(1)) durchgeführt wurden. Es bestünde folglich die Möglichkeit, dass die Substanzen der EGFR-Inhibitor-Serie inhibitorisch auf andere BET-Proteine oder auf die BD2 wirken. Allerdings sind die BDs der BET-Proteine stark konserviert und nur wenige selektive Inhibitoren bekannt.

Außerdem muss bedacht werden, dass es sich bei den analysierten Inhibitoren um Prototypen handelt, welche noch hinsichtlich Potenz und Selektivität optimiert werden müssen. Optimierte Inhibitoren würden in Zukunft einen detaillierteren Einblick in die zellulären Zusammenhänge erlauben.

Anschließend wurden verschiedene Arten von PPARγ-Modulatoren getestet.

III.3.2.7. Biochemische Charakterisierung von PPARγ-Modulatoren

In der Therapie wurden bisher überwiegend PPARγ-Agonisten (voll oder partial) eingesetzt (*DrugBank* 2018). Jedoch gibt es ebenfalls PPARγ-Antagonisten. Eine Unterscheidung nach der Art der Modulation wurde zuvor nicht vorgenommen. Aus diesem Grund sollte nun untersucht werden, ob je nach Art der PPARγ-Modulation ein anderer Effekt auf BRD4 beobachtet werden kann.

Substanz-ID	Struktur	Ø Restaktivität ± σ in % bei einer Konzentration von 10 μM	30 μM	50 μM	Art
GW9662		96,5 ± 6,02	101,9 ± 2,29	99,1 ± 1,64	Antagonist
FH535		98,6 ± 3,97	93,6 ± 2,72	90,5 ± 3,05	Antagonist
Basalazid		102,7 ± 3,60	101,0 ± 4,67	99,4 ± 0,09	Agonist (Mesalazin Prodrug)
Ibuprofen		100,0 ± 1,38	97,0 ± 4,83	97,2 ± 2,67	Partial-agonist
Mesalazin		84,9 ± 5,64	64,3 ± 3,82	53,7 ± 1,02	Agonist

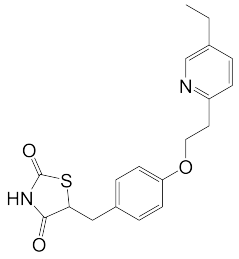
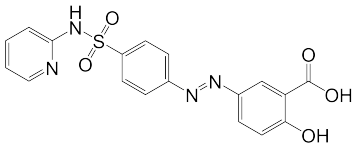
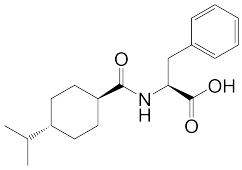
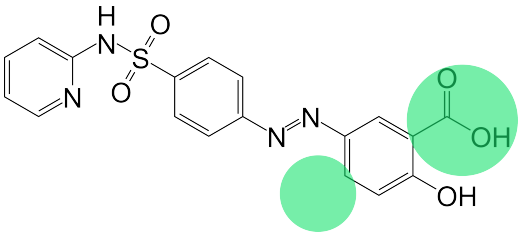
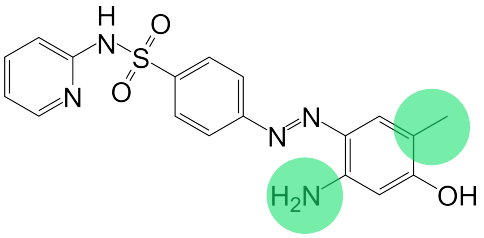
Substanz-ID	Struktur	Ø Restaktivität ± σ in % bei einer Konzentration von 10 μM	30 μM	50 μM	Art
Pioglitazon		98,8 ± 5,98	99,3 ± 5,07	92,7 ± 1,02	Agonist
Sulfasalazin		100,2 ± 3,69†	99,8 ± 3,52 †	100,6 ± 4,48†	Agonist (Mesalazin Prodrug)
Nateglinid		95,6 ± 6,56	92,8 ± 0,95	89,7 ± 6,48	Agonist

Tabelle III.3.5.: Ergebnisse der PPARγ-Modulatoren bezüglich ihrer inhibitorischen Wirkung auf BRD4. †: Eigenfluoreszenz und Interferenz mit HTRF-Assaysystem.

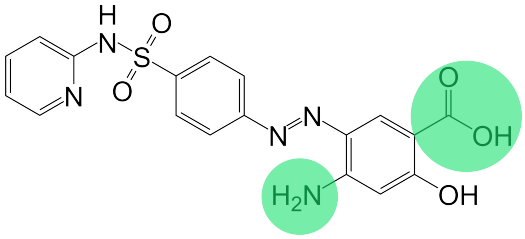
Alle getesteten Antagonisten (GW9662 und FH535) sowie die meisten Agonisten zeigten keine dosisabhängige BRD4-Inhibition. Ein Effekt kann aufgrund der vorliegenden Daten nur für Mesalazin angenommen werden und ein in Abhängigkeit von der Wirkung auf PPARγ unterschiedlicher Effekt auf BRD4 ist mit den vorliegenden Daten nicht ermittelbar. Da es sich bei vielen der getesteten Substanzen um auf dem Markt befindliche Arzneimittel bzw. Toolsubstanzen handelt, kann vermutet werden, dass während ihrer Entwicklung auf Selektivität hin optimiert wurde. Dies könnte die fehlende Aktivität auf BRD4 begründen. Nichtsdestotrotz ist das Ergebnis für Sulfasalazin überraschend. Eine sehr ähnliche Verbindung (MS436) ist als potenter BRD4-Inhibitor bekannt (siehe Abbildung III.3.9) (Zhang *et al.* 2013). Aus diesem Grund sollte der Sachverhalt genauer geprüft werden. Dazu wurde eine intermediäre Verbindung zwischen Sulfasalazin und MS436, genannt SSLH01 (Abbildung III.3.9c), durch Saskia Spitzer und Dr. Luis Bering synthetisiert. Sie besitzt einerseits die Carboxygruppe von Sulfasalazin und andererseits die primäre aromatische Aminogruppe von MS436. Leider zeigten Sulfasalazin und SSLH01 eine Interferenz mit dem HTRF-Assaysystem und der K_d-Wert liegt für beide Verbindungen im orthogonalen MST-Assay (Tabelle III.3.2) oberhalb von 100 μM.



(a) Sulfsalazin (**22**)



(b) MS436
K_i = 30-50 nM



(c) SSLH01 (**21**)

Abbildung III.3.9.: Strukturen von Sulfasalazin (PPARγ-Agonist), MS436 (BRD4-Inhibitor) und SSLH01 (Hybrid-Verbindung).

Sulfasalazin besteht aus 5-Aminosalicylsäure (Mesalazin, **12**) und Sulfapyridin (**9**) und wird enzymatisch unter physiologischen Bedingungen in diese Bestandteile gespalten, womit es ein Prodrug von Mesalazin ist. Für Mesalaszin kann eine geringe dosisabhängige BRD4-Inhibition vermutet werden (siehe Tabelle III.3.5) zudem wurde ein Shift der Schmelztemperatur von 0,78 °C im TSA erhalten (siehe Tabelle III.3.2). Außerdem konnte für Sulfapyridin ein K_d-Wert von 7,3 μM mittels MST bestimmt werden (siehe Tabelle III.3.2). Daüberhinaus gelang die Kokristallisation von Sulfaslazin (**22**) und SSLH01 (**21**) mit BRD4. Die Ergebnisse werden im Abschnitt III.3.2.9 diskutiert.

III.3.2.8. Ergebnisse der orthogonalen Assays

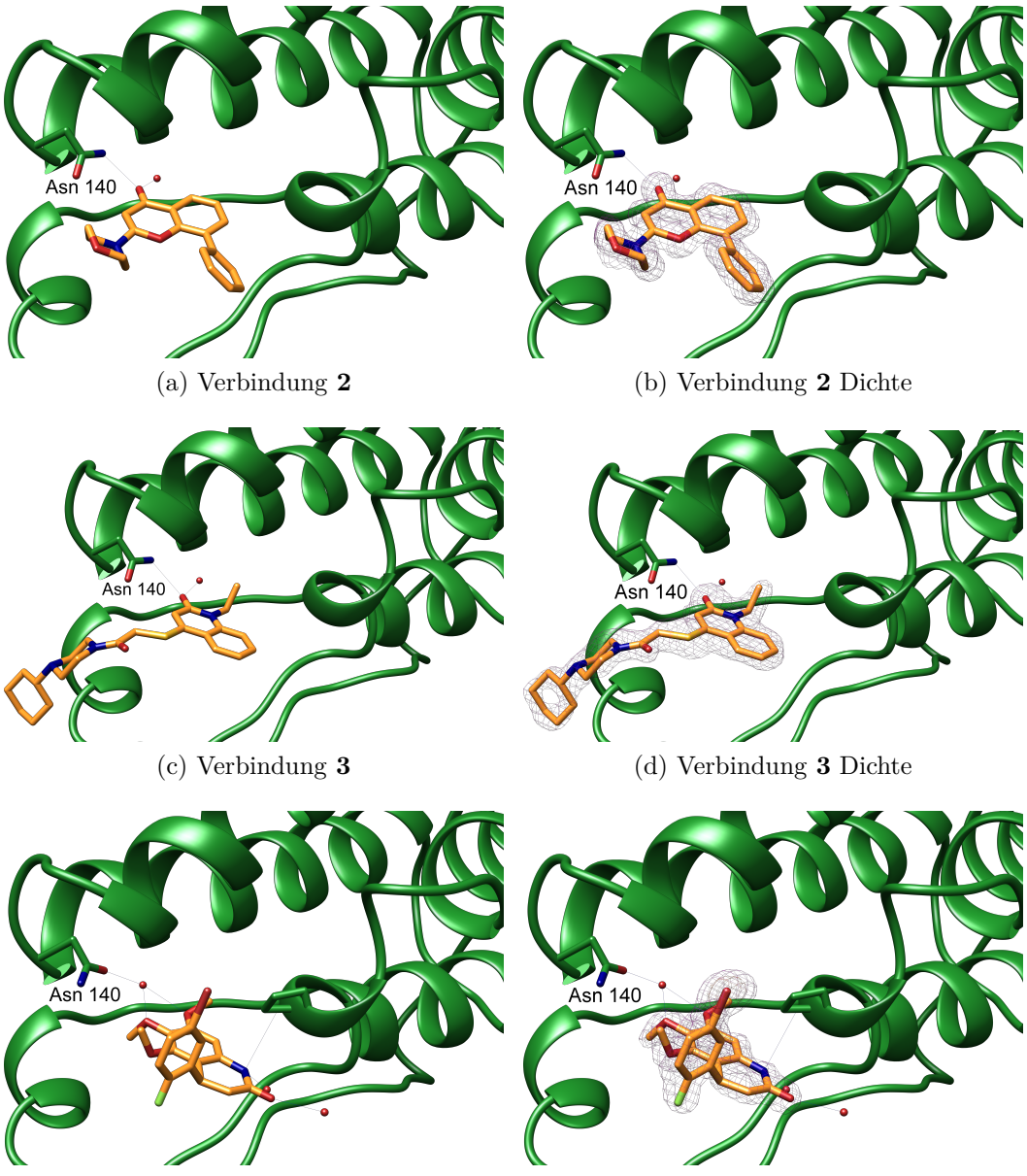
Als orthogonale Assays zur Verifizierung der durch den HTRF-Assay gefundenen Hits wurden TSA und MST von Dr. Jette Pretzel verwendet. Mithilfe des TSA wurde die Veränderung der Schmelztemperatur (auch Denaturierungstemperatur) für das Holoprotein in Bezug auf das Apoprotein (BRD4) ermittelt. Zwölf der 22 getesteten Verbindungen erhöhten die Schmelztemperatur signifikant (siehe Tabelle III.3.2). Dies weist auf eine Stabilisierung des Proteins durch die Bindung des Analyten hin. Allerdings sind die ermittelten Veränderungen in der Schmelztemperatur (Shift) im Bezug auf die in der Literatur angegebenen Shifts von bekannten BRD4-Inhibitoren wie z. B. JQ1 vergleichsweise gering. Die *chemical probe* JQ1 wurde als Kontrolle verwendet und konnte einen mit der Literatur vergleichbaren Shift erzielen (Filippakopoulos *et al.* 2010). Ursache für die vergleichsweise geringen Shifts könnte die geringe Potenz (im unteren mikromolaren Bereich) der untersuchten Inhibitoren im Vergleich zu der hohen Potenz (im nanomolaren Bereich) von JQ1 sein. Mithilfe des MST-Assays wurden die Bindungsaffinitäten interessanter initialer Hits bestimmt. Die K_d-Werte befinden sich in der Tabelle III.3.2. Neun der elf initialen Hits, die mittels MST getestet wurden, konnten bestätigt (K_d im unteren mikromolaren Bereich) werden. Von den zwöf im HTRF-Assay als inaktiv ($IC_{50} > 10$ μM) bestimmten Verbindungen konnten neun per MST als inaktiv bestätigt werden. Es wurden per MST drei falsch positive Hits (**6**: Docking, **10** und **17**: APSC PPARγ) und ein falsch negativer Hit (**9**: Sulfapyridin) identifiziert. Verbindung **6** induziert eine starke Aggregation von BRD4, die Destabilisierung zeigt sich im TSA durch den signifikanten negativen Shift (ΔTm = - 4,55 °C). Eine initiale Verringerung der Fluoreszenz konnte bei der Verbindung **17** festgestellt werden. Vermutlich liegt ebenfalls eine Destabilisierung von BRD4 vor.

Für die Verbindungen **15** und **16** wurden hohe K_d-Werte von 87,43 ± 0,97 μM bzw. 104,70 ± 3,71 μM ermittelt. Bei solch hohen Konzentrationen können vermehrt Effekte auftreten, welche zu einer Beeinflussung des Assaysystems oder einer nicht spezifischen Wechselwirkung mit dem untersuchten Protein führen können. Zu diesen Effekten zählen transiente Präzipitation, Lichtstreuung und Aggregation (Jadhav *et al.* 2010). Zur Verbesserung der Zuverlässigkeit des Assaystems wird der Zusatz von Detergenzien, in diesem Fall wurde Polysorbat 20 (Tween 20) verwendet, empfohlen. Trotzdem wäre zur Verifizierung der Inhibition und im Hinblick auf eine spätere Verwendung in zellulären Untersuchungen eine Analyse des Selektivitätsprofils ratsam. Bei der in zellulären Assays einzusetzenden, für den Effekt auf BRD4 benötigten Konzentration scheint wahrscheinlich, dass weitere Proteine beeinflusst werden würden und demzufolge nicht der Effekt auf BRD4 im zellulären Assay vordergründig und für den beobachteten Phänotyp maßgeblich verantwortlich sein könnte.

III.3.2.9. Kokristallstrukturen von BRD4

Zur weiteren Verifizierung der Hits konnten zwei mittels APSC gefundene Hits (**2** und **3**), ein aus dem Docking-Ansatz stammender Hit (**4**) sowie Sulfaslsazin und SSLH01 mit BRD4 erfolgreich kokristallisiert werden. Die Statistiken zur Datensammlung und Verfeinerung befinden sich im Anhang in der Tabelle B.7. Der Bindemodus der Verbindungen **2**, **3** und **4** entspricht dem in den meisten bekannten BRD4-Inhibitor-Komplexstrukturen anzutreffenden Bindemodus. Die entsprechenden drei Kristallstrukturen wurden von Dr. Jette Pretzel gelöst. Zu dem Bindemodus gehört eine Wasserstoffbrückenbindung zum Asparagin 140 und Interaktionen mit konservierten Wassermolekülen in der Bindetasche von BRD4(1) (Abbildung III.3.10). Die Verbindung **2** ($K_d = 4{,}08 \pm 0{,}33$ μM, $IC_{50} = 5{,}1 \pm 0{,}59$ μM) ist besonders interessant, da sie das durch das APSC identifizierte, privilegierte Chromon-Scaffold enthält. Die Verbindung ist ein bekannter Kinase-Inhibitor (LY294002, Phosphoinositid-3-Kinase p110α: $IC_{50} = 0{,}5$ μM, $K_i = 6$ μM (Bento *et al.* 2014)) und wurde bereits als BRD4-Inhibitor in der Literatur beschrieben ($IC_{50} = 7{,}30$ μM) (Dittmann *et al.* 2014). Die Verbindung **4** ($K_d = 1{,}64 \pm 0{,}03$ μM, $IC_{50} = 0{,}58 \pm 0{,}08$ μM) ist der potenteste in dieser Arbeit im HTRF-Assay gefundene BRD4-Inhibitor. Interessant ist, dass in dieser Komplexstruktur das Asparagin 140 in einer anderen Konformation vorliegt. Dies liegt darin begründet, dass die Wechselwirkung zwischen Asparagin 140 und Ligand nicht direkt über eine Carbonyl-Gruppe des Liganden (**2**, **3**) oder die Sulfoxid-Gruppe von Dimethylsulfoxid (DMSO) wie in den Komplexstrukturen mit Sulfasalazin und SSLH01 erfolgt, sondern über ein Wassermolekül vermittelt wird. Eine bekannte BRD4-Komplexstruktur mit ähnlichem Liganden und ebenfalls einer Wasser-vermittelten Wasserstoffbrückenbindung weist ebenfalls die veränderte Konformation des Asparagin 140 auf (PDB-ID: 4O78).

Sulfasalazin und SSLH01 zeigen im Vergleich zu den anderen Inhbitoren dieser Analyse und den meisten bisher kokristallisierten Inhibitoren einen anderen Bindemodus (Abbildung III.3.11). Interessant ist, dass der Bindemodus für Sulfasalazin und SSLH01 ebenfalls von dem Bindemodus für MS436 (PDB-ID: 4NUD, (Zhang *et al.* 2013)), einem Analogon mit inhibitorischer Wirkung auf BRD4 (Abbildung III.3.9), abweicht. Der Bindemodus ähnelt eher dem des bivalenten Inhibitors biBET (Waring *et al.* 2016), welcher mit zwei Bromodomänen (BDs) gleichzeitig interagiert und eine hochpotente (im picomolareren Bereich) *chemical probe* (*Chemical Probes* 2018) ist. Die Bindemodi von MS436, biBET und Sulfasalazin in BRD4 sind in Abbildung III.3.12 zusammenfassend und im Anhang in der Abbildung B.9 einzeln dargestellt. BiBET interagiert mit Asparagin 140, während Sulfaslazin und SSLH01 nicht direkt mit Asparagin 140 interagieren. Anstelle des Inhibitors interagiert DMSO, welches als Lösungsmittel verwendet wurde, mit Asparagin 140. Die beobachtete Lokalisation von DMSO in der BRD4-Bindetasche wurde bereits beschrieben (Lolli und Battistutta 2013). Eine Überlagerung der gelösten BRD4-Sulfasalazin-Komplexstruktur mit der beschriebenen BRD4-DMSO-Komplexstruktur befindet sich in der Abbildung B.9 im Anhang.



(a) Verbindung **2**

(b) Verbindung **2** Dichte

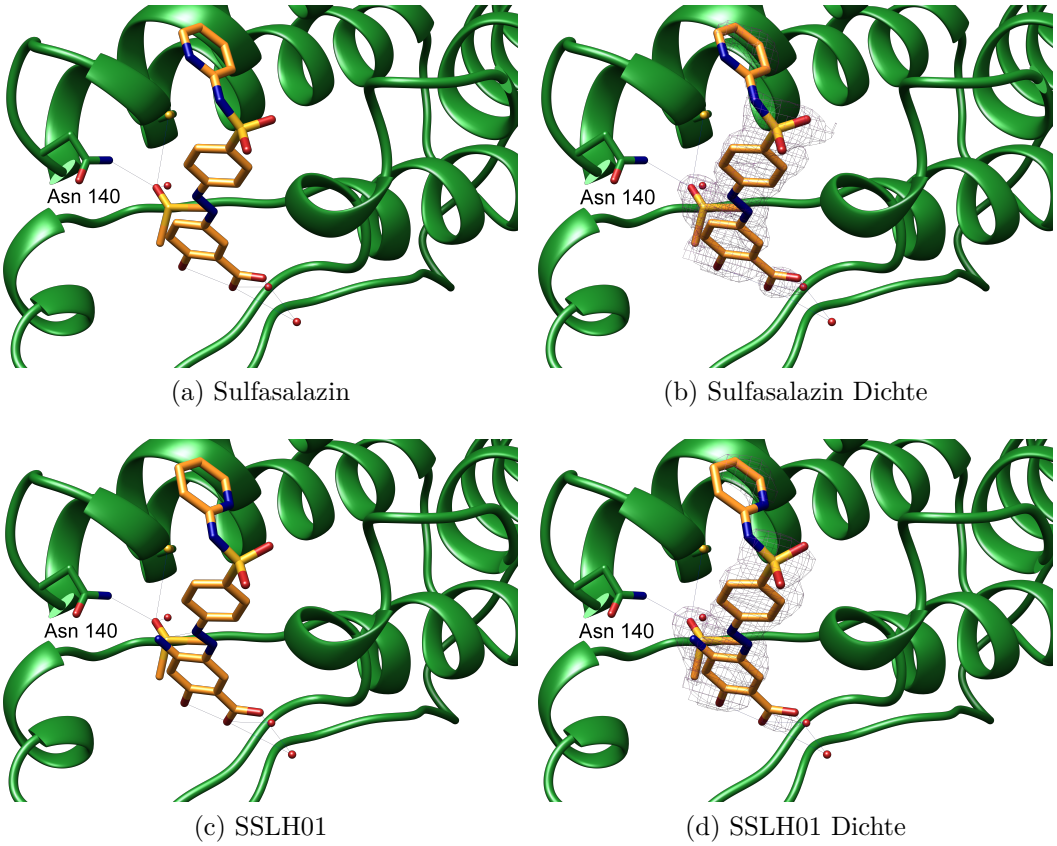
(c) Verbindung **3**

(d) Verbindung **3** Dichte

(e) Verbindung **4**

(f) Verbindung **4** Dichte

Abbildung III.3.10.: Komplexstrukturen von BRD4 mit den Verbindungen **2** (Auflösung 1,31 Å), **3** (1,61 Å) und **4** (1,62 Å). Der Ligand ist orange eingefärbt. In der rechten Spalte wird die $2F_o$-F_c Elektronendichte des Liganden konturiert bei 1 RMSD gezeigt.



(a) Sulfasalazin (b) Sulfasalazin Dichte

(c) SSLH01 (d) SSLH01 Dichte

Abbildung III.3.11.: Komplexstrukturen von BRD4 mit Sulfasalazin (Auflösung 1,48 Å) und SSLH01 (1,91 Å). Der Ligand ist orange eingefärbt. In der rechten Spalte wird die $2F_o$-F_c Elektronendichte des Liganden konturiert bei 1 RMSD gezeigt.

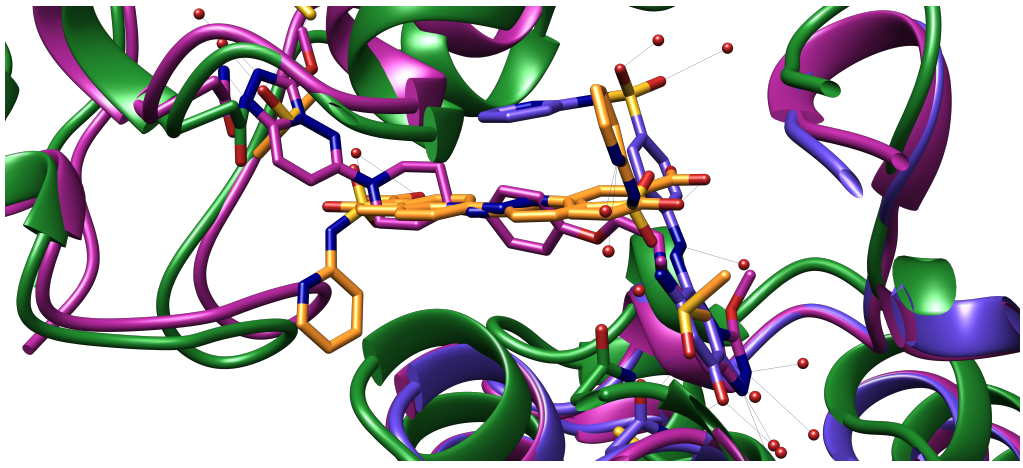


Abbildung III.3.12.: Alignment der BRD4-Sulfasalazin-Komplexstruktur (als Dimer basierend auf Kristallkontakten) (grün) mit BRD4-MS436- und BRD4-biBET-Komplexstrukturen (PDB-IDs: 4NUD (blau) und 5AD3 (lila)). Sulfasalazin ist orange eingefärbt.

III.3.2.10. Bedeutung der entdeckten Ähnlichkeit von BRD4 und PPARγ: Resistenzen und neue therapeutische Strategien

Mit der Entdeckung der Ähnlichkeit zwischen BRD4 und PPARγ sowohl auf der Protein- als auch auf der Ligand-Ebene, stellte sich die Frage nach einer physiologischen Beziehung der beiden Proteine und deren therapeutischen Nutzbarmachung. Eine intensive Literaturrecherche offenbarte ein komplexes physiologisches Zusammenspiel von BRD4 und PPARγ. Eine Zusammenfassung der gefundenen Zusammenhänge befindet sich unter dem Abschnitt II.2.3. Auffallend ist, dass beide Proteine an einer Vielzahl gemeinsamer Krankheiten (z. B. Krebs, inflammatorischen und kardiovaskuläre Erkrankungen) beteiligt sind, welche insbesondere im Zusammenhang mit oxidativem Stress stehen.

BETi-Resistenzen, welche sowohl intrinsisch als auch erworben sein können, sind ein sehr aktuelles Forschungsgebiet und einige Kombinationstherapien zur Überwindung der Resistenz wurden bereits aufgezeigt (Pawar *et al.* 2018; Wu *et al.* 2018). Eine Liste mit resistenten Zelllinien befindet sich in Rathert *et al.* (2015). Zwei mögliche Kombinationsstrategien sollen im Folgenden näher betrachtet werden. Einerseits die bereits in der Literatur diskutierte Kombination von BETi und Kinase-Inhibitoren und andererseits eine neue Kombinationsstrategie von BETi und PPARγ-Agonisten, welche sich aus den Ergebnissen der vorliegenden Arbeit ergibt.

Für duale Kinase-BET-Inhibitoren wurden kürzlich positive Auswirkungen auf die Tumortherapie diskutiert (Carlino und Rastelli 2016) und ihr Wert insbesondere für die Entwicklung einer Anti-Resistenzstrategie wurde demonstriert (Kurimchak *et al.* 2016; Ember *et al.* 2017).

Insbesondere zwischen CDK9 und BRD4 besteht eine enge Beziehung. CDK9 bildet mit einem Cyclin (z. B. Cyclin T1) den positiven Transkriptionselongationsfaktor P-TEFβ. BRD4 rekrutiert P-TEFβ und beeinflusst so die Transkription über die Regulierung der Aktivität der RNA-Polymerase II (Itzen *et al.* 2014). Eine Koinhibition von CDK9 und BRD4 wurde in mehreren Arbeiten beschrieben (Brägelmann *et al.* 2017; Pawar *et al.* 2018; Lu *et al.* 2015). Ebenfalls konnten synergistische Effekte für die Kombination von BRD4- und CDK2-Inhibitoren beobachtet werden (Bolin *et al.* 2018). Dies ist interessant, da CDK2 durch das APSC als mit BRD4 über ähnliche Liganden verbundenes Protein identifiziert wurde. Außerdem teilen CDK9 und BRD4 ebenfalls ähnliche Liganden wie z. B. Flavopiridol (Abbildung III.3.2a). Eine weitere Strategie gegen BETi-Resistenzen ist die Kombination von einem MEK-Inhibitor (RAS-Signalweg) und einem BETi (Wyce *et al.* 2018). Da EGFR ebenfalls einen Einfluss auf den RAS-Signalweg besitzt, sind die in dem Abschnitt III.3.2.6 aufgezeigten dualen EGFR-BET-Inhibitoren RL078 und RL129 ebenfalls potentiell von hoher therapeutischer Relevanz zur Bekämpfung von resistenten Tumoren. Zudem wurde EGFR als möglicher Bestandteil der BETi- und Rezeptortyrosinkinaseinhibitor-Resistenzentwicklung identifiziert (Kurimchak *et al.* 2016; Kleczko und Heasley 2018). Interessant ist, dass bei der Anpassung der Tumorzelle an eine Rezeptortyrosinkinaseinhibition, z. B. EGFR-Inhbition, der NFκB-Signalweg aktiviert wird und somit das Überleben des Tumors gefördert wird (Kleczko und Heasley 2018).

Daraus ergibt sich ein therapeutisches Potential für die Kombination aus einem EGFR-Inhibitor und einerseits einem Inhibitor, welcher die Anpassung der Tumorzelle durch Transkriptionsinhibition, z. B. durch einen BETi, verhindert und/oder andererseits einem Inhibitor des NFκB-Signalwegs, wie z. B. einem PPARγ-Agonisten (siehe auch Abbildungen II.0.6 und III.3.13). Zudem bietet ein zielgerichteter polypharmakologischer Ansatz das Potential dem Teufelskreis der „Ko-Evolution“ von Proteinmutationen und Inhibitorgenerationen zu entkommen, der sich als bedeutende Herausforderung insbesondere der zielgerichteten Therapie herauskristallisiert hat. Für das Design von EGFR-Inhibitoren formulierten Engel, Lategahn und Rauh 2016 den Bedarf einer Alternative zu einer (einfachen) zielgerichteten Therapie. Diese könnte in dem hier präsentierten zielgerichteten und spezifischen polypharmakologischen (mehrfach zielgerichteten) Ansatz liegen. Tumorzellen sind mitunter sehr flexible Systeme, die häufig schnell auf nur eine Störung (z. B. einen Rezeptortyrosinkinaseinhibitor) reagieren und sich anpassen können. Werden sie jedoch durch mehrere gezielte Störungen aus ihrem Gleichgewicht gebracht, ist eine länger anhaltende Wachstumshemmung oder Apoptose denkbar und eine Resistenzbildung möglicherweise verzögert bis verhindert (Kleczko und Heasley 2018). Des Weiteren kann eben diese Flexibilität auch therapeutisch in Kombinations-Ansätzen ausgenutzt werden (Ishay-Ronen *et al.* 2019). Dazu sind insbesondere polypharmakologische Verbindungen der Klassen 1 (zwei Moleküle verbunden durch einen Linker wie z. B. PROTACs) und 4 (hoch affine Moleküle mit spezifischer Aktivität auf definierten nicht verwandten Proteinen) nach Meyers *et al.* (2018) geeignet.

Für BETs und PPARγ könnte das Zusammenspiel insbesondere zur Erhaltung eines Tumors ebenfalls therapeutisch bedeutsam sein. Eine wichtige Rolle beim Erhalt des Tumors spielt in vielen Tumorzelllinien das Protoonkogen MYC. Wird die MYC-Expression durch einen Inhibitor eines MYC-regulierenden Signalweges gehemmt, erfolgt häufig eine kompensatorische Hochregulierung eines anderen MYC-regulierenden Signalweges und es entsteht eine Resistenz gegenüber dem eingesetzten Inhibitor. Einige an der MYC-Expression beteiligte Signalwege und Proteine sind unter anderem: BRD4 (Delmore *et al.* 2011), Wnt/β-Catenin (Engelke und Chinnaiyan 2015), Hedgehog (Kumar *et al.* 2015) und NFκB (Wu *et al.* 2018). Es existiert ein Crosstalk zwischen dem Wnt/β-Catenin- und dem Hedgehog-Signalweg und die Expression von GLI1 und GLI2 (Bestandteile des Hedgehog-Signalweges) kann durch β-Catenin reguliert werden (Song *et al.* 2015; Kumar *et al.* 2015). Wnt/β-Catenin und NFκB wurden als wichtige Signalwege in resistenten Zellen identifiziert (Engelke und Chinnaiyan 2015; Wu *et al.* 2018; Rathert *et al.* 2015). Sie erhalten eine hohe MYC-Expression trotz BRD4-Inhibition. Sowohl BRD4 als auch PPARγ kontrollieren die Expression von MYC. Dabei erhöht BRD4 die MYC-Expression, während PPARγ durch die Hemmung der beiden wichtigen Signalwege NFκB und Wnt/β-Catenin die MYC-Expression senkt (Abbildung III.3.13).

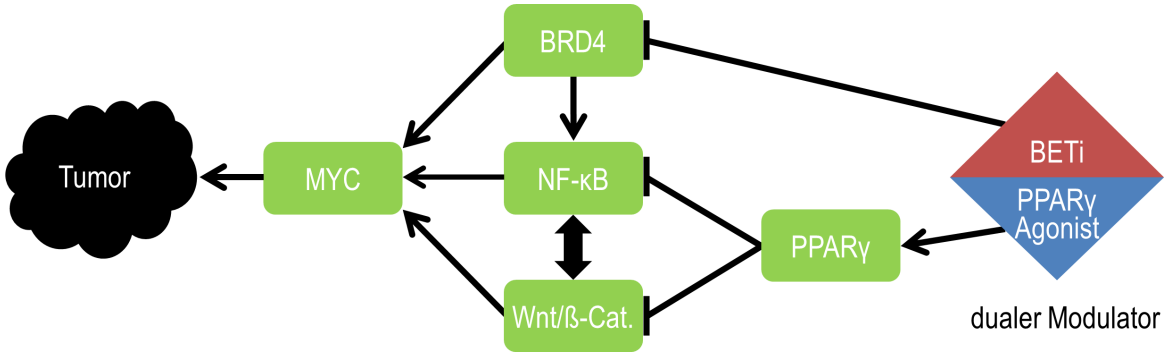


Abbildung III.3.13.: Zusammenspiel von BRD4 und PPARγ zur Aufrechterhaltung des Tumors sowie Darstellung des Wirkmechanismus eines dualen Modulators.

Demzufolge wird durch die gleichzeitige Inhibition der MYC-Expression durch eine BETi vermittelte BET-Inhibition sowie durch eine Inhibition der NFκB- und Wnt/β-Catenin Signalwege, vermittelt durch eine Aktivierung von PPARγ mithilfe eines PPARγ-Agonisten, eine kompensatorische Erhöhung der MYC-Expression durch einen dieser Mechanismen unterbunden (Abbildung III.3.13). Diese Ergebnisse liefern die Basis für den therapeutischen Nutzen eines dualen BRD4-PPARγ-Modulators. Dabei kann ein dualer BRD4-PPARγ-Modulator auf zweifache Weise genutzt werden, einerseits zur Prävention einer Resistenzentstehung und andererseits zur Behandlung einer bestehenden Resistenz. Zusammen mit den Beobachtungen zu Kinase-Inhibitoren wäre außerdem ein Tripel-Modulator Kinase-BET-PPARγ oder eine entsprechende Kombinationstherapie zur zweifachen Bekämpfung von Resistenzen (Behandlung und Prävention) denkbar. Mögliche Startpunkte für einen solchen Tripel-Modulator wären die beiden EGFR-BRD4-Inhibitoren RL078 und RL129 sowie Luteolin. Das Wirkprofil der EGFR-BRD4-Inhibitoren müsste auf PPARγ erweitert werden. Zudem wäre eine Optimierung der Potenz und der Selektivität erforderlich. Das Selektivitätsprofil von Luteolin müsste auf die Inhibition von wichtigen Kinasen, z. B. CDK2 oder CDK9 und BRD4 sowie die PPARγ-Aktivierung hin optimiert werden. Trotz der guten Fortschritte in der Erforschung von möglichen Ansätzen zur Bekämpfung von Resistenzen und der Vielzahl an aufgezeigten Kombinationsmöglichkeiten, darf die Heterogenität und Plastizität der Tumore nicht vernachlässigt werden. Aus diesem Grund ist neben der weiteren Erforschung von therapeutischen Strategien ebenfalls die Identifizierung von zuverlässigen Biomarkern zur Vorhersage der Wirksamkeit und von potentiellen Umgehungsstrategien einer solchen Strategie für die im Patienten gefundenen Genotypen des Tumors essentiell, um eine geeignete Therapieoption auszuwählen.

III.3.3. Zusammenfassung

In diesem Kapitel wurde die Scaffold-Ebene der Ähnlichkeit unter Einbeziehung von Proteinzielstrukturen betrachtet. Dies folgte aus den vorhergehenden Analysen der Bioähnlichkeit von niedermolekularen Verbindungen, die die Wichtigkeit der Einbeziehung der Proteine, welche für die biologische Wirkung verantwortlich sind, herausgestellt haben (siehe Kapitel III.2). Insbesondere sollte der Frage nachgegangen werden, ob eine Ähnlichkeit zweier niedermolekularer Verbindungen basierend auf einem gemeinsamen (privilegierten) Scaffold eine Ähnlichkeit der adressierten Proteine impliziert. Dazu wurde ein Workflow, genannt SPARROW, entwickelt, welcher die Scaffold-Ebene mit der Protein-Ebene verbindet. Die Anwendung von SPARROW auf die DrugBank lieferte das Chromon-Scaffold als ein interessantes privilegiertes Scaffold. In weiterführenden Analysen des Chromon-Scaffolds und der dazugehörigen Proteinzielstrukturen BRD4, PPARγ, CDK2, ESR2 und NR1H4 konnte das Chromon-Scaffold als erweitertes privilegiertes Scaffold identifiziert werden. Dies bedeutet, dass basierend auf der Ähnlichkeit der Liganden auf der Grundlage des gemeinsamen Chromon-Scaffolds eine Ähnlichkeit zwischen einigen der fünf genannten Proteine in einem Bindetaschenvergleich mittels IsoMIF gefunden werden konnte. Daraufhin wurden Verbindungen aus einer internen Bibliothek (COMAS-Bibliothek), welche ähnlich zu den Liganden der Proteine PPARγ, CDK2, ESR2 oder NR1H4 sind, bezüglich ihrer inhibitorischen Wirksamkeit gegenüber BRD4 getestet. Zum Vergleich wurden ebenfalls Verbindungen der COMAS-Bibliothek, welche in einem Docking-Ansatz ermittelt wurden, getestet. Der Docking-Ansatz und das APSC erzielen eine ähnlich hohe Hitrate von ca. 5 %. Jedoch unterscheiden sich die gefundenen Hits in ihrem chemischen Raum, womit die Methoden komplementär sind. Die hohe Hitrate von 14,34 % der PPARγ ähnlichen Verbindungen ist beeindruckend. Einige der ausgewählten Hits konnten in orthogonalen Assays (TSA und MST) validiert werden. Von fünf Verbindungen gelang die Kokristallisation mit BRD4. Darunter ist Sulfasalazin, ein auf dem Markt befindlicher PPARγ-Agonist. Interessanterweise zeigen Sulfasalazin und das Derivat SSLH01 einen von den anderen drei kokristallisierten Inhibitoren distinkten Bindemodus. Unter den Hits wurde auch ein Inhibitor aus einer Serie von EGFR-Varianten-Inhibitoren identifiziert. In einer anschließenden Analyse von weiteren Inhibitoren der Serie konnten zwei weitere duale EGFR-BRD4-Inhibitoren identifiziert werden. Die vorliegenden zellulären Daten wurden analysiert und Unterschiede in den häufigen Mutationen von gegenüber dualen EGFR-BRD4-Inhibitoren sensitiven und weniger sensitiven (insensitven) Zelllinien identifiziert. Für genauere Aussagen sollten diese Analysen auf der Expressionsebene wiederholt werden. Insgesamt ist die Entdeckung der Ähnlichkeit von BRD4 und PPARγ, welche sich einerseits durch eine hohe, deutlich über dem Zufall liegende Bindetaschenähnlichkeit und andererseits durch eine Vielzahl von ähnlichen Liganden (hohe Hitrate) bemerkbar macht, überraschend und eindrucksvoll. Sie ist insbesondere im Hinblick auf neue Therapieoptionen wertvoll.

Interessant ist, dass BRD4 Liganden mit Kinasen und PPARγ teilt und eine Kombination von BRD4- und Kinase-Inhibitoren und/oder PPARγ-Agonisten therapeutisch sinnvoll scheint.

In Zukunft sollte das APSC breiter angewendet werden und mithilfe von SPARROW zu neuen erweiterten privilegierten Scaffolds führen, um somit die Wirkstoffentwicklung zu fördern. Zum Beispiel könnte die Ligand-Ähnlichkeit des APSC, die im vorgestellten ersten Ansatz auf gemeinsamen Scaffolds basiert, durch zusätzliche Elemente, die z. B. gemeinsame tolerierte bioisostere Austausche oder eine ähnliche dreidimensionale Gestalt (Shape) erfassen, erweitert werden.

Eine erste Untersuchung dieser Ähnlichkeitsebenen von niedermolekularen Verbindungen befindet sich im nächsten Kapitel. Polypharmakologische Ansätze scheinen zur Behandlung und Prävention von Resistenzen wirkungsvoll und sollten weiter, zusammen mit potentiellen Biomarkern, analysiert werden. Insbesondere müssen Methoden zum rationalen Design von polypharmakologischen Wirkstoffen entwickelt werden. Dazu liefert die vorliegende Arbeit mit dem APSC und SPARROW gute Ansatzpunkte. Diese resultierten bereits in der Entdeckung neuer BRD4-Inhibitoren sowie der Beziehung von BRD4 und PPARγ, die therapeutisch zur Bekämpfung von Resistenzen in der Tumortherapie genutzt werden kann.

Quintessenz:

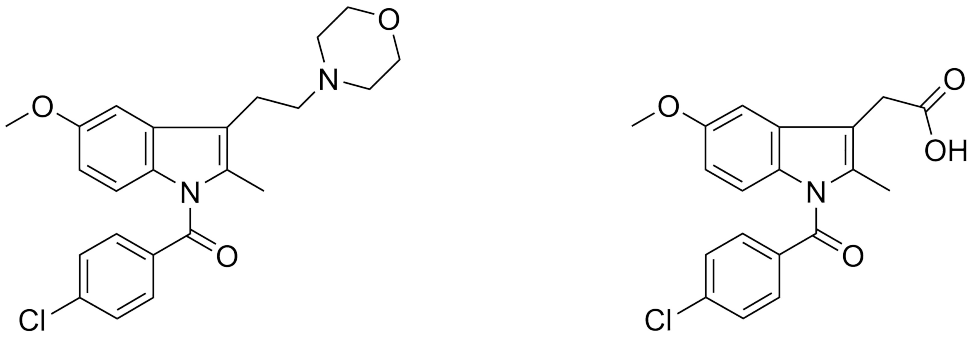
- erfolgreiche Entwicklung und Validierung eines Workflows zur Identifizierung möglicher *advanced privileged scaffolds*, genannt SPARROW
- Bestätigung des entwickelten *advanced privileged scaffold concepts* (APSC), da eine Ligand-Ähnlichkeit auf eine Protein-Ähnlichkeit übertragen werden konnte
- neue Inhibitoren für BRD4 wurden identifiziert, darunter duale BRD4-PPARγ- und BRD4-EGFR-Modulatoren
- Beziehung zwischen BRD4 und PPARγ wurde entdeckt, diese kann im Kampf gegen Resistenzen durch duale BRD4-PPARγ-Modulatoren genutzt werden

III.4. Die Bioisoster- und Shape-Ebene der Ähnlichkeit

III.4.1. Einleitung

Dieses Kapitel beschäftigt sich vornehmlich mit der Ähnlichkeit von Atomgruppen. Im Gegensatz zu der Pharmakophor-Ebene werden jedoch keine Atome zusammengefasst und es entstehen keine reduzierten Graphen. Die Suche nach chemisch distinkten Atomen oder Atomgruppen, welche dennoch ähnliche Eigenschaften aufweisen, hat eine lange Tradition und das rationale Erkennen von Ähnlichkeit ist ein langersehntes Ziel vieler Wissenschaftler. Dabei wurde auf verschiedene Eigenschaften fokussiert. So wurden im Laufe der Zeit viele Definitionen von Isosteren und Bioisosteren aufgestellt und erweitert (Meanwell 2011). Eine für die medizinische Chemie bedeutende Definition lieferte Thornber: „Bioisosteres are groups or molecules which have chemical and physical similarities producing broadly similar biological properties“ (Thornber 1979). Um eine eindeutige und für den Computer verarbeitbare Interpretation von „broadly similar“ zu gewährleisten, wurden in dieser Arbeit Atome, Atomgruppen und Moleküle, deren Bioaktivität sich um weniger als eine Größenordnung unterscheidet, als bioähnlich definiert. Außerdem wurde eine chemische oder physikalische Ähnlichkeit nicht überprüft, sondern auf die Bioähnlichkeit fokussiert, da das Ziel der vorliegenden Arbeit ist, die Bioähnlichkeit zu untersuchen. Zudem ist es anzunehmen und gewünscht, dass sich einige Eigenschaften ändern. Dies kann dazu genutzt werden, um z. B. die Zellpermeabilität bei gleichbleibender Aktivität zu erhöhen. Um bioisostere Gruppen identifizieren zu können, wurden zunächst *matched molecular pair*s (MMPs) mittels Pipeline Pilot in einer Selektion von ChEMBL-Molekülen identifiziert. Ein MMP besteht aus zwei Molekülen, die sich nur in einer Gruppe unterscheiden. Folglich kann der Effekt der differenzierenden Gruppe z. B. auf die Bioaktivität analysiert werden. Ist der Effekt auf die Bioaktivität kleiner als eine Größenordnung, wird in dieser Arbeit von einer bioisosteren Gruppe gesprochen. Ein Beispiel für ein MMP befindet sich in Abbildung III.4.1. Als differenzierende Gruppe wurde in diesem Beispiel ein Essigsäure-Rest III.4.1b und entsprechend ein 4-Ethylmorpholin-Rest III.4.1a gefunden.

Eng mit der Bioisoster-Ebene verbunden, ist die Shape-Ebene. Eine bestimmte räumliche Anordnung von spezifischen funktionellen Gruppen (Shape) ist oft eine Voraussetzung für eine Interaktion der niedermolekularen Verbindung mit dem Zielprotein.



(a) CHEMBL72771 (b) CHEMBL6

Abbildung III.4.1.: Beispiel für ein MMP: CHEMBL72771 (a) und CHEMBL6 (Indometacin) (b).

Folglich ist die Shape-Ähnlichkeit eines bioisosteren Paares meist hoch. Jedoch kann z. B. durch die Flexibilität des Proteins auch ein anderer Shape zu einer ähnlichen Bioaktivität führen.

Trotzdem sollen die beiden Ebenen in diesem Kapitel teilweise getrennt betrachtet werden, da sie verschiedene Voraussetzungen mit sich führen. So ist für die Betrachtung der Bioisoster-Ebene ein Zugang zu Bioaktivitätsdaten essentiell, während die Shape-Ebene auf eine dreidimensionale Struktur angewiesen ist. Obwohl das Scaffold einen maßgeblichen Einfluss auf den Shape des Moleküls besitzt, ist die Umkehrung weniger eindeutig. So lässt sich vom Shape nicht wieder auf das zugrundeliegende Scaffold zurückschließen. In der Regel ist ein bestimmter Shape nicht ausreichend, allerdings notwendig für die Interaktion mit dem Zielprotein. In dieser Arbeit wurden zwei Methoden zur Analyse bzw. Bewertung der Shape-Ähnlichkeit angewendet: (i) *normalized principal moments of inertia ratio* (NPR) (Sauer und Schwarz 2003) und (ii) FastROCS (OpenEye) (*FastROCS Toolkit* 2018). NPR ist unabhängig von einer Referenz-Struktur und FastROCS bezieht das Alignment von physikochemischen Eigenschaften (genannt *Color Features*) mit ein. Eine genauere Beschreibung der beiden Verfahren befindet sich unter Abschnitt IV.2.20.

Ein bioisosterer Austausch wird oft während der Optimierung eines Arzneistoffkandidaten in Betracht gezogen, um zum Beispiel Parameter aus den Bereichen *absorption, distribution, metabolism, excretion, toxicology* (Resorption, Verteilung, Metabolismus, Ausscheidung, Toxikologie) (ADMET) zu verbessern. Sowohl die Bioisoster- als auch die Shape-Ebene (sowie ebenso die Pharmakophor-Ebene) kann zum Scaffold-Hop, einer Veränderung des Grundgerüstes, verwendet werden. Ziel des *Scaffold-Hoppings* kann das Verlassen eines patentgeschützen chemischen Raumes oder die Entdeckung von innovativen, aktiven und zum Beispiel mit weniger problematischen Gruppen belasteten Strukturen sein.

III.4.2. Ergebnisse und Diskussion

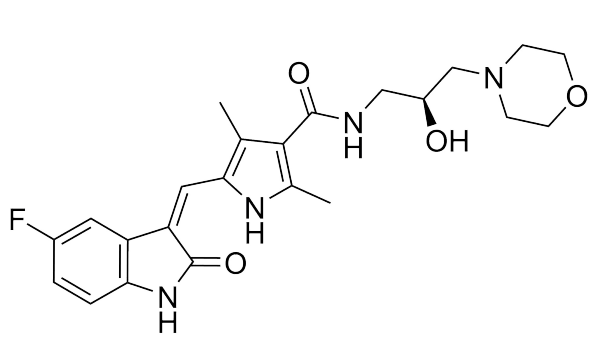
III.4.2.1. Bioisoster-Ebene

Aus der ChEMBL-Selektion wurden mehr als 31 Millionen MMPs gewonnen. Daraus konnten 17.851.377 nicht einzigartige bioisostere Paare generiert werden. Mithilfe von Filterungsschritten, welche durch Python-Skripte realisiert wurden, ergaben sich mehr als eine Millionen einzigartige und 93.282 verlässliche bioisostere Paare. Verlässlich ist ein bioisosteres Paar, wenn es mindesten 20 mal gefunden wurde und mindestens 75 % der Austausche zu keiner Änderung der Bioaktivität führten. Da die vollständige Darstellung aller identifizierten Austausche den Rahmen der Arbeit sprengen würde, sollen die Resultate mithilfe von Beispielen veranschaulicht werden. Die im Folgenden aufgeführten Bioaktivitätsdaten entstammen der ChEMBL (Bento *et al.* 2014).

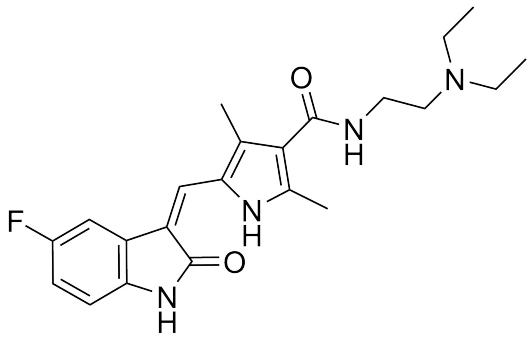
Ein Austausch von einer azyklischen Gruppe mit einer zyklischen Gruppe, wie im MMP-Beispiel in Abbildung III.4.1 gezeigt, konnte ebenfalls häufig in den verlässlichen bioisosteren Paaren gefunden werden. Einen Bioisosterenaustausch ebenfalls mit einem Morpholin-Rest zeigt die Abbildung III.4.2. Der in den Abbildungen III.4.2a und III.4.2b gezeigte Austausch erfolgt zwischen einer azyklischen (Triethlyamin-Rest, Abbildung III.4.2b) und einer zyklischen Gruppe (Morpholin-Rest, Abbildung III.4.2a). Dahingegen ist in den Abbildungen III.4.2c und III.4.2d ein Austausch von zwei zyklischen Gruppen (Pyrrolidin und Morpholin) abgebildet. Interessanterweise gehören die Arzneistoffe Sunitinib (Struktur siehe III.4.2b) und Gefitinib (Struktur siehe III.4.2d) beide der Klasse der Rezeptortyrosinkinaseinhibitoren an.

Oft konnte ein Austausch eines Kohlenstoffatoms gegen eine größere Gruppe gefunden werden. Zwei entsprechende Beispiele sind in der Abbildung III.4.3 abgebildet. Beim ersten Beispiel (Abbildungen III.4.3a und III.4.3b) handelt es sich um einen Spezialfall des Austausches einer azyklischen Gruppe gegen eine zyklische Gruppe (3,5-Dimethylbenzen-Rest, Abbildung III.4.3a), da die azyklische Gruppe nur aus einem Atom, hier dem Kohlenstoffatom, besteht. Im zweiten Beispiel sind beide Gruppen azyklisch (Kohlenstoffatom, Abbildung III.4.3c und Methylformiat-Rest, Abbildung III.4.3d).

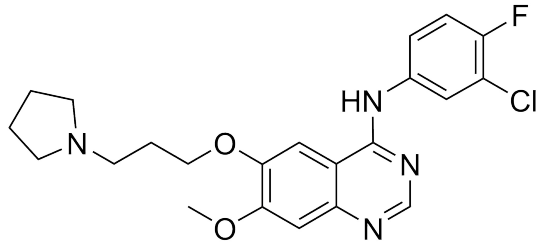
Analog zum Austausch eines Kohlenstoffatoms kann ebenfalls der Austausch eines Wasserstoffatoms ein verlässliches bioisosteres Paar darstellen. Zwei Beispiele dafür sind in der Abbildung III.4.4 veranschaulicht. Die im oberen Beispiel dargestellten Strukturen der Dopamintransporter-Inhibitoren weisen ein bioisosteres Paar auf, welches aus einer zyklischen (3-Picolin-Rest, Abbildung III.4.4a) und einer azyklischen (Wasserstoffatom, Abbildung III.4.4b) Gruppe besteht. Beim bioisosteren Paar mit Ketoprofen erfolgt ein Austausch einer azyklische Gruppe (Ethylnitrat-Rest, Abbildung III.4.4c) gegen eine andere azyklische Gruppe (Wasserstoffatom, Abbildung III.4.4d).



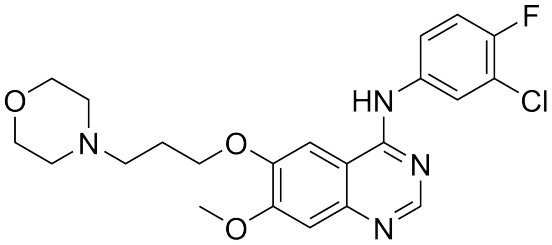
(a) CHEMBL1721885
K_d = 4,7 nM



(b) CHEMBL535, Sunitinib
K_d = 1,8 nM

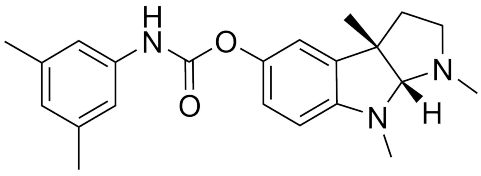


(c) CHEMBL299672
IC_{50} = 71 nM

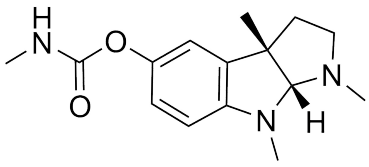


(d) CHEMBL939, Gefitinib
IC_{50} = 23 nM

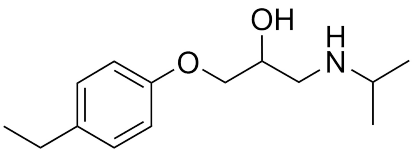
Abbildung III.4.2.: Zwei Beispiele für bioisostere Paare mit Morpholin-Rest: CHEMBL1721885 (a) und CHEMBL535 (Sunitinib) (b) beide inhibieren den *vascular endothelial growth factor receptor 1* (Vaskulärer endothelialer Wachstumsfaktor-Rezeptor 1) (VEGFR1) (UniProt-ID: P17948) sowie CHEMBL299672 (c) und CHEMBL939 (Gefitinib) (d) beide inhibieren den EGFR (UniProt-ID: P00533).



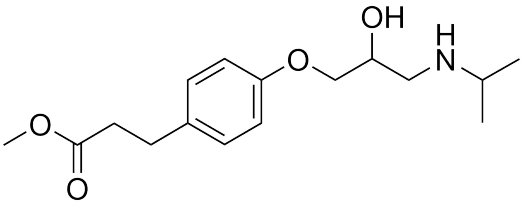
(a) CHEMBL131210
IC_{50} = 78 nM



(b) CHEMBL94, Physostigmin
IC_{50} = 56,7 nM

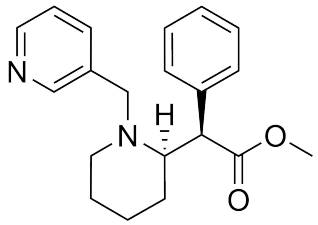


(c) CHEMBL143467
K_d = 63,1 nM

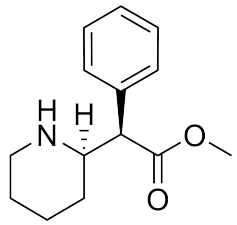


(d) CHEMBL768, Esmolol
K_d = 100 nM

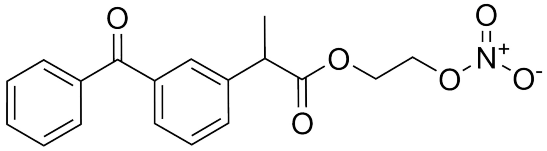
Abbildung III.4.3.: Zwei Beispiele für bioisostere Paare mit einem Kohlenstoffatom: CHEMBL131210 (a) und CHEMBL94 (Physostigmin) (b) beide inhibieren die AChE (UniProt-ID: P22303) sowie CHEMBL143467 (d) und CHEMBL768 (Esmolol) (d), welche beide Antagonisten der β_1- und β_2-Adrenozeptoren sind (UniProt-IDs: P10608 und P18090). Die angegebenen K_d-Werte wurden für β_1-Adrenozeptoren des Hausmeerschweins ermittelt (UniProt-ID: B0FL73).



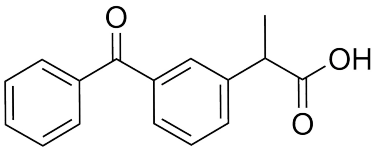
(a) CHEMBL1255014
IC_{50} = 173 nM



(b) CHEMBL827, Dexmethylphenidat
IC_{50} = 83 nM

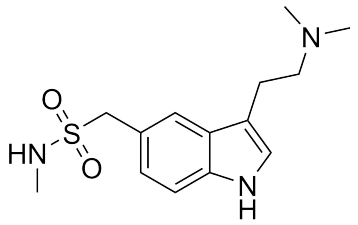


(c) CHEMBL444396
COX1: IC_{50} = 450 nM
COX2: IC_{50} = 770 nM

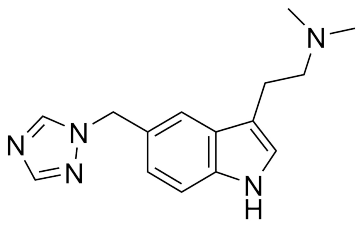


(d) CHEMBL571, Ketoprofen
COX1: IC_{50} = 330 nM
COX2: IC_{50} = 690 nM

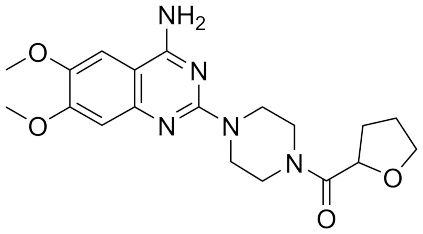
Abbildung III.4.4.: Zwei Beispiele für bioisostere Paare mit einem Wasserstoffatom: CHEMBL1255014 (a) und CHEMBL827 (Dexmethylphenidat) (b) beide inhibieren den Dopamintransporter (UniProt-ID: P23977) sowie CHEMBL444396 (d) und CHEMBL571 (Ketoprofen) (d), welche beide Inhibitoren der Cyclooxygenase-1 (COX-1) und COX-2 sind (UniProt-IDs: P23219 und P35354).



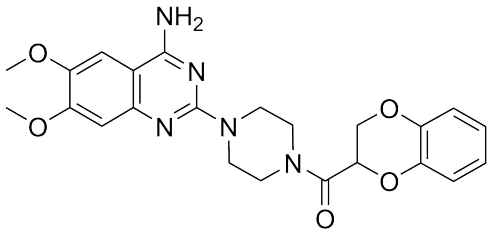
(a) CHEMBL128, Sumatriptan
5-HT$_{1B}$: IC$_{50}$ = 16 nM
5-HT$_{1D}$: IC$_{50}$ = 5 nM



(b) CHEMBL905, Rizatriptan
5-HT$_{1B}$: IC$_{50}$ = 41 nM
5-HT$_{1D}$: IC$_{50}$ = 11 nM



(c) CHEMBL611, Terazosin
K_d = 36,31 nM



(d) CHEMBL707, Doxazosin
K_d = 25,7 nM

Abbildung III.4.5.: Zwei Beispiele für bioisostere Paare mit zyklischen Gruppen: die beiden Triptane CHEMBL128 (Sumatriptan) (a) und CHEMBL905 (Rizatriptan) (b) (UniProt-IDs: P28222 und P28221) sowie CHEMBL611 (Terazosin) (c) und CHEMBL707 (Doxazosin) (d) beide sind α_{1A}-Adrenozeptor-Antagonisten (UniProt-ID: P35348).

Wie bereits am Beispiel der Rezeptortyrosinkinaseinhibitoren gezeigt, ist ebenfalls ein Austausch einer azyklischen Gruppe gegen eine zyklische Gruppe sowie der Austausch zwischen zwei zyklischen Gruppen häufig unter Beibehaltung einer ähnlichen Bioaktivität möglich. Zwei bioisostere Paare, welche solche Austausche zeigen, werden beispielhaft anhand von Arzneistoffen in Abbildung III.4.5 gezeigt.

Der bioisostere Austausch eines Carbonsäure-Restes durch ein Sulfonamid oder ein Tetrazol ist bekannt (Ballatore, Huryn und Smith 2013). In den ermittelten verlässlichen bioisosteren Paaren konnte ein Austausch eines Sulfonamid-Restes (Sumatriptan, Abbildung III.4.5a) gegen einen 1,2,4-Triazol-Rest (Rizatriptan, Abbildung III.4.5b) gefunden werden. Ein bioisosterer Austausch eines Carbonsäure-Restes gegen einen Triazol-Rest wurde ebenfalls bereits in der Literatur beschrieben (Bonandi *et al.* 2017; Song *et al.* 2013). Allerdings ist entweder der Triazol-Rest nicht über ein Stickstoffatom sondern über ein Kohlenstoffatom mit dem restlichen Molekül verbunden (Song *et al.* 2013) oder es handelt sich um 1,2,3-Triazole (Bonandi *et al.* 2017). In den Abbildungen III.4.5c und III.4.5d wird das bioisostere Paar Tetrahydrofuran (Terazosin) und Benzodioxan (Doxazosin) gezeigt. Abschließend soll darauf hingewiesen werden, dass ein bioisosteres Paar, auch wenn es verlässlich ist, nicht zwangsläufig für jedes Protein anwendbar sein muss.

III.4.2.2. Shape-Ebene

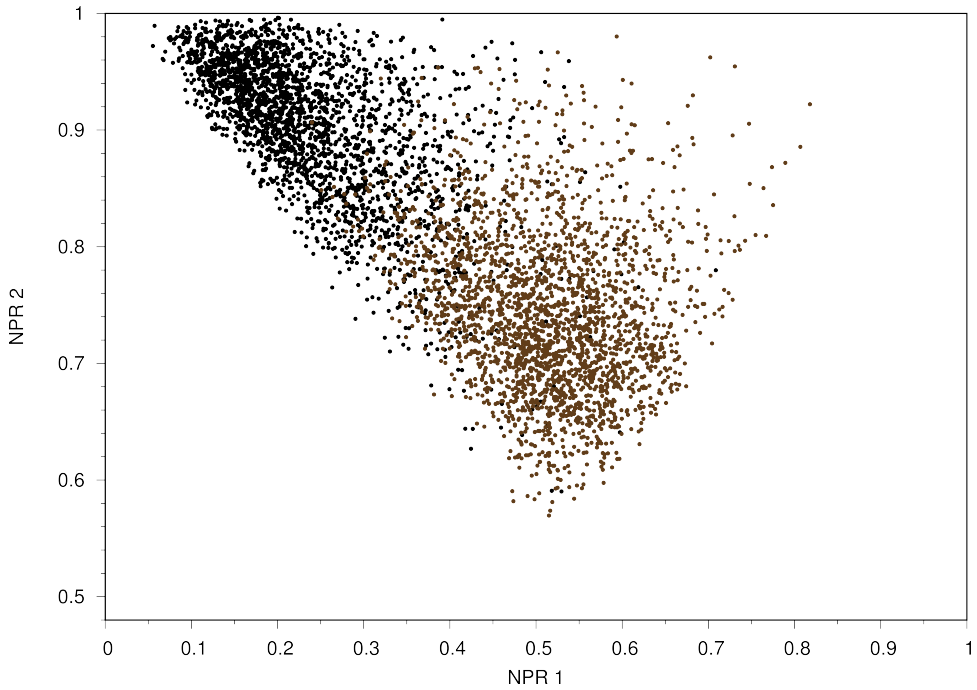
Die Analyse der Shape-Ebene erfolgte sowohl mithilfe der absoluten Methode der NPRs als auch mittels einer relativen Methode unter Verwendung von *Cambridge Structural Database* (CSD) Fragmenten und FastROCS.
Ein nicht zu vernachlässigender Punkt, welcher die Shape-Ebene betrifft, ist, dass die bioaktive Konformation der niedermolekularen Verbindung selten bekannt ist und folglich meistens vorhergesagt werden muss. Einerseits ist diese Vorhersage mit Fehlern behaftet und andererseits ist sie rechenintensiv und damit nur bedingt für große Molekül-Datenbanken geeignet. Ein möglicher Lösungsansatz wäre für häufig vorkommende Fragmente, z. B. Benzenring, die Konformationen und NPR-Werte zu berechnen. Anschließend müssten die zu untersuchenden Moleküle fragmentiert werden und ein Gesamt-NPR-Wert approximiert werden. Sicherlich wird die für das Molekül vorherrschende Konformation weder stets die biologisch aktive Konformation sein noch wird sie zwangsläufig mit der Konformation des Fragmentes ohne weitere Substituenten identisch sein. Für die CSD-basierte Shape-Ähnlichkeit konnte bereits auf experimentell ermittelte Konformationen zurückgegriffen werden. Dennoch müssen diese nicht mit der biologisch aktiven Konformation übereinstimmen und die Konformationen der extrahierten Grundgerüste können ebenfalls durch die Art und Anzahl der Substituenten beeinflusst werden. Die Implementierung der Shape-Ebene in den vorhandenen BBPMCS-Algorithmus gestaltet sich wesentlich komplexer als die Implementierung der Bioisoster-Ebene, die auf bioisosteren Paaren basiert, welche nur über eine Bindung mit dem restlichen Molekül verbunden sind. Einerseits kann die Anzahl und Position der Substituenten in einem Shape-Paar variieren und andererseits ist die Orientierung der Shape-ähnlichen Fragmente relevant. Bezüglich der Position der Substituenten wurde in dieser Arbeit der Lösungsweg über den Erhalt des am Grundgerüst adjazenten Substituentenatoms (*Substituenten-Dummies*) erarbeitet. Eine variierende Anzahl von Substituenten spielt vornehmlich bei der Enumeration von Shape-ähnlichen Molekülen und nicht für Clustering- oder Such-Algorithmen, welche auf synthetisierte Moleküle angewendet werden, eine Rolle. Durch Verwendung der *Substituenten-Dummies* kann hier ebenfalls ein Lösungsansatz aufgezeigt werden. Die Berücksichtigung der Orientierung der Fragmente stellt eine wirkliche Herausforderung dar, da dies ein Alignment der Fragmente voraussetzt. Der BBPMCS-Algorithmus erzeugt ein Mapping der Atome und erstellt somit ein Alignment der verglichenen Graphen. Allerdings ist dies nur unter bestimmten Bedingungen möglich, z. B. bei identischen Graphen. Für bioisostere oder Shape-ähnliche Gruppen ist aktuell kein Alignment mittels BBPMCS umsetzbar. Deshalb wurden diese beiden Ebenen, wie in diesem Kapitel dargelegt, separat modelliert. Der Shape eines Moleküls müsste zusätzlich zum Molekülgraphen als Graph repräsentiert werden, um ein Shape-Mapping per Mapping der Shape-Graphen zu erlauben.

NPR

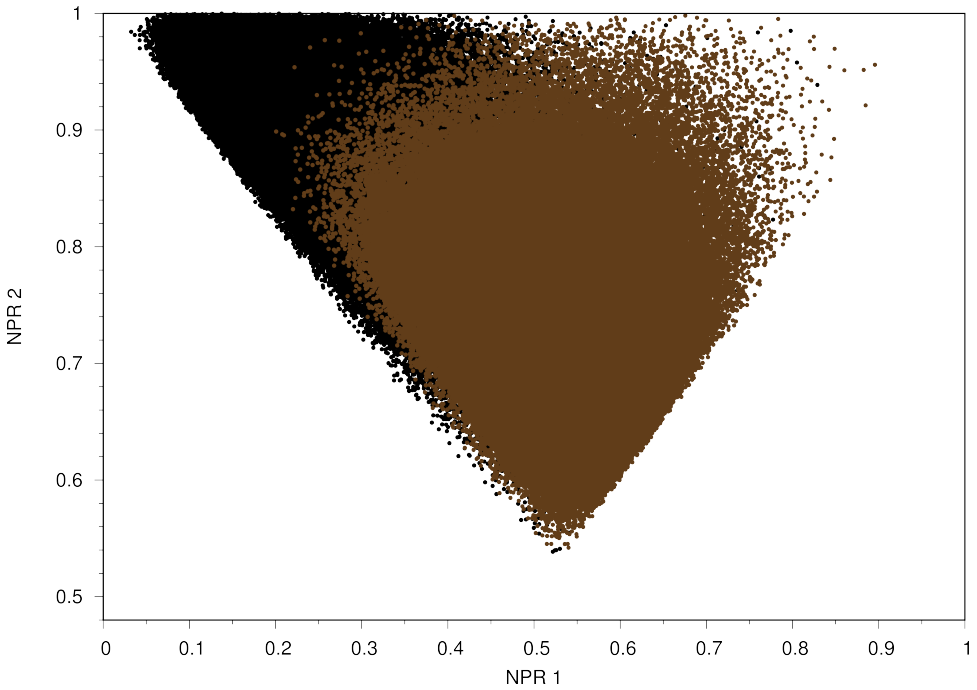
Ein Ziel bei der Erstellung von Molekülbibliotheken ist, diese möglichst divers zu gestalten. Die Diversitätsanalyse ist eng mit der Ähnlichkeitsanalyse verwandt. So wurde die Diversität verschiedener *in silico* synthetisierter DELs mittels NPR-Analyse und einer *principal component analysis* (Hauptkomponentenanalyse) (PCA) der Deskriptoren (i) Anzahl Wasserstoffbrückendonatoren, (ii) Anzahl Wasserstoffbrückenakzeptoren, (iii) Anzahl rotierbarer Bindungen, (iv) SlogP, (v) TPSA und (vi) Molekülmasse untersucht. Die Berechnung der Deskriptoren und der PCA erfolgte mithilfe von MOE auf den, wie unter Abschnitt IV.2.20 beschrieben, vorbereiteten dreidimensionalen Strukturen der Moleküle. Im Rahmen dieses Kapitels sollen die Ergebnisse der NPR-basierten Diversitätsanalyse vorgestellt werden. Die Abbildung III.4.6 zeigt die Shape-Verteilung der *DNA ecoded libraries* (DNA-kodierten Bibliotheken) (DELs) basierend auf den Reaktionen 7 und 8 (Reaktionsgleichungen siehe Abschnitt IV.2.3). Bei der Betrachtung des Shape-Raumes der kleinen Bibliotheken (2.500 Produkte, Abbildung III.4.6a) ist der Unterschied besonders markant. Während die Mitglieder der Bibliothek 7 eher einen linearen stabförmigen Shape (linke, obere Ecke) aufweisen, zeichnen sich die Mitglieder der Bibliothek 8 durch einen vornehmlich scheibenförmigen Shape (untere Ecke) aus und die Bibliothek 8 scheint inhärent einen größeren Shape-Raum abzudecken. Es ist anhand des Vergleiches der Shape-Räume der kleinen (Abbildung III.4.6a) mit denen der großen (250.000 Produkte, Abbildung III.4.6b) Bibliotheken deutlich erkennbar, dass die Erhöhung der Eduktanzahl den Shape-Raum der Produkte einer Reaktion erweitern kann, sofern diese rational durch Diversitätsanalysen ausgewählt wurden.

CSD-FastROCS

Aus der CSD wurden alle kohlenstoffhaltigen Verbindungen gewonnen und per Python-Skript von Dr. Hitesh Patel wurden die zentralen Grundgerüste mit und ohne die Beibehaltung eines Atoms je Substituent extrahiert. Im Ansatz ohne Beibehaltung von Substituentenatomen wurden 15.615 verschiedene Scaffolds erhalten, unter Beibehaltung von *Substituenten-Dummies* konnten 39.414 unterschiedliche Scaffolds extrahiert werden. Jedes der so generierten Scaffolds wurde mit jedem anderen aus demselben Ansatz (ohne bzw. mit *Substituenten-Dummies*) mithilfe von FastROCS verglichen und eine entsprechende Distanzmatrix gewonnen. Außerdem wurden *Substitutionspfade* (siehe Abbildung III.4.7) und *Äquivalenzklassen* (siehe Abbildungen III.4.9, III.4.10 und III.4.11) basierend auf einem TanimotoCombo-Ähnlichkeitsscore von mindestens 1,7 ermittelt. Ein Substitutionspfad bedeutet, dass man theoretisch Substruktur A durch Substruktur C austauschen kann, wenn eine Substitutionsregel von A nach B und von B nach C existiert. In einer Äquivalenzklasse werden nur solche Substitutionen zusammengefasst, die in sich geschlossen sind, dies bedeutet, dass es von jeder Substruktur in der Äquivalenzklasse eine Substitutionsregel zu jeder anderen Substruktur in derselben Äquivalenzklasse gibt und keine weiteren Substitutionsregeln existieren.



(a) kleine Bibliotheken



(b) große Bibliotheken

Abbildung III.4.6.: Beispiel für einen Shape-Vergleich mittels NPR. Verglichen wurde der Shape-Raum der Produkte der DEL Reaktionen 7 (schwarz) und 8 (braun) für eine kleine Bibliothek (a) sowie eine große Bibliothek (b).

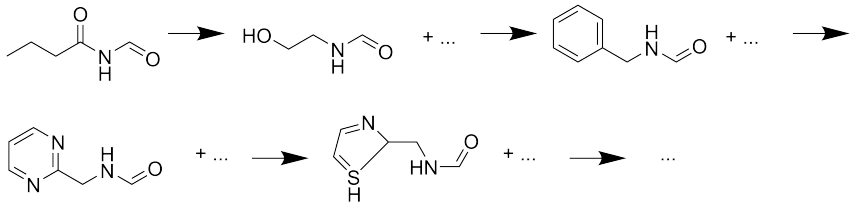


Abbildung III.4.7.: Beispiel für einen Substitutionspfad.

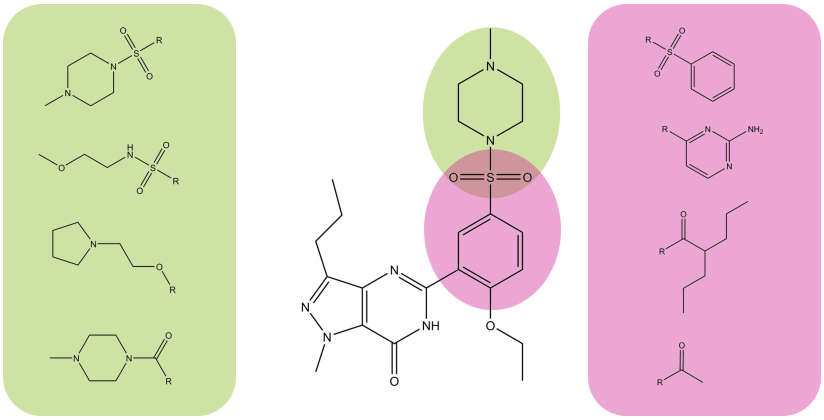


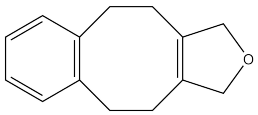
Abbildung III.4.8.: Beispiel für überlappende austauschbare Substrukturen (hier bioisostere Paare) für Sildenafil.

III.4.2.3. Herausforderungen und deren mögliche Lösungen

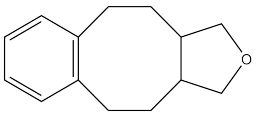
Die Bioisoster- und die Shape-Ebene ermöglichen Substitutionspfade. Ein Beispiel für einen Substitutionspfad befindet sich in Abbildung III.4.7.

Bezüglich der Substitutionspfade stellt sich die Frage, wie oft eine oder verschiedene Substitutionsregeln angewendet werden dürfen, ohne eine maßgebliche Änderung der Bioaktivität hervorzurufen. Das bedeutet wie lang der Substitutionspfad sein darf, ohne dass die Bioähnlichkeit der Startsubstruktur zur Zielsubstruktur verloren geht. Eine Antwort müsste vermutlich empirisch ermittelt werden. Neben der Ermittlung der Anzahl möglicher Substitutionen stellt auch die Bestimmung der Reihenfolge dieser eine Herausforderung dar. Die Reihenfolge ist bei überlappenden austauschbaren Substrukturen von Bedeutung (siehe Abbildung III.4.8). Zur Lösung des Problems könnten alle Substitutionspfade in allen Reihenfolgen getestet werden. Eine solche Vorgehensweise hätte jedoch negative Auswirkungen auf die Laufzeit.

Eine Alternative zu Substitutionspfaden wären Äquivalenzklassen. Werden statt Substitutionspfaden Äquivalenzklassen verwendet, so sind diese im Falle der Shape-Ebene stark beschränkt auf vorwiegend große starre Grundgerüste. Drei Beispiele solcher Äquivalenzklassen befinden sich in den Abbildungen III.4.9, III.4.10 und III.4.11.

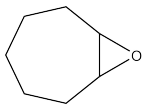


(a) CHEMBL71046

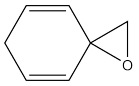


(b) CHEMBL72675

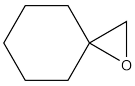
Abbildung III.4.9.: Beispiel für eine Äquivalenzklasse mit drei Ringsystemen.



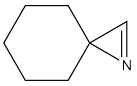
(a) CHEMBL71046



(b) CHEMBL72675



(c) CHEMBL71046

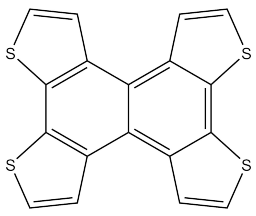


(d) CHEMBL72675

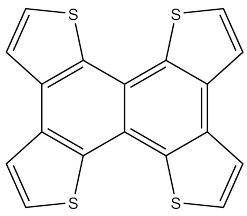
Abbildung III.4.10.: Beispiel für eine Äquivalenzklasse mit zwei Ringsystemen.

Eine weitere Möglichkeit bestünde in der Vorberechnung aller möglichen Substitutionen. Für die CSD-basierten Substrukturen der Shape-Ebene wurde dieser Lösungsansatz durch die Generierung von Distanzmatrizen bereits verfolgt. Allerdings ist dieser nur anwendbar, wenn Austausche ausschließlich zwischen vordefinierten Substrukturen erlaubt sind. Erfolgt die Ermittlung der Ähnlichkeit hingegen über ein virtuelles Intermediat, welches einen vordefinierten Substrukturaustausch darstellt und die größte Ähnlichkeit zu dem zu bewertenden Molekül besitzt, muss eine zweistufige Evaluation über einerseits die Distanzmatrix und andererseits über eine weitere Ebene der Ähnlichkeit, z. B. der fconv-Atomtypengruppen-Ebene, erfolgen.

In diesem Kapitel wurden nur solche bioisosteren Paare beschrieben, die exakt über eine Bindung mit dem restlichen Teil des Moleküls verknüpft sind. Diese Einschränkung hat den Vorteil, dass einige Probleme nicht beachtetet werden müssen, wie z. B. die Orientierung des bioisosteren Paares zum restlichen Molekül.



(a) CHEMBL71046



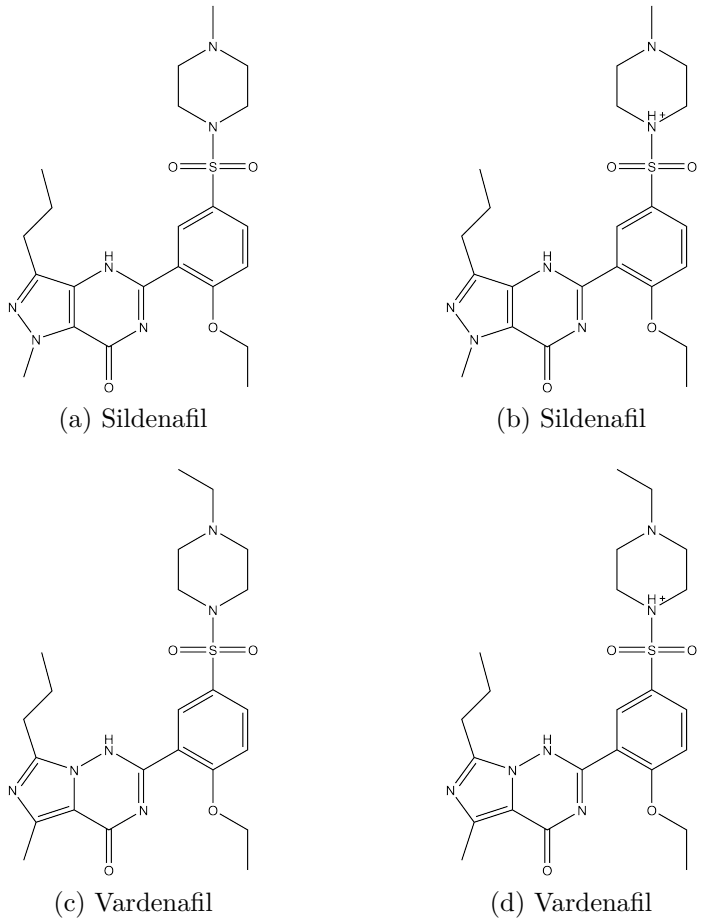
(b) CHEMBL72675

Abbildung III.4.11.: Beispiel für eine Äquivalenzklasse mit sechs Ringsystemen.

Um eventuelle negative Auswirkungen eines Austausches zu kompensieren, können die Substituenten und das Substitutionsmuster des Austauschpaares variieren (Thornber 1979). Damit wären zwei Moleküle mit verschiedenem Substitutionsmuster unter Umständen ähnlicher als zwei Moleküle mit gleichem Substitutionsmuster. Dies in einen Algorithmus zu fassen, gestaltet sich jedoch äußerst schwierig.

Ein bisher vernachlässigter jedoch für das Hervorrufen eines biologischen Effektes essentieller Aspekt sind die Zustände, in denen sich ein Molekül befinden kann. Damit sind insbesondere verschiedene Protonierungszustände (Beispiel siehe Abbildung III.4.12) und Tautomerformen (Beispiel siehe Abbildung III.4.13) gemeint. Zudem ist ebenfalls die Anzahl solcher Zustände von Interesse. Je flexibler ein Molekül ist, mit desto mehr verschiedenen Proteine kann es potentiell interagieren. Auf der anderen Seite kann die Interaktion eines flexibleren Moleküles mit einem höheren Energieaufwand verbunden sein. Um den Aspekt der Zustände zu berücksichtigen, könnten diese in einem Vorbereitungsschritt enumeriert werden. Dies würde sich jedoch zweifach negativ auf die Laufzeit auswirken, da einerseits die Enumeration zusätzliche Zeit beansprucht und andererseits sich die Anzahl der Eingabegraphen (zu untersuchende Moleküle) für z. B. Clustering- oder Suchalgorithmen, vervielfacht. Eine weitere Option bestünde in der Veränderung der Atomtypendefinitionen, sodass die Atomtypen nicht von der Anzahl adjazenter Wasserstoffatome abhängen. Zusätzlich würde eine Repräsentation ohne Wasserstoffatome gewählt werden. Hier wäre im Allgemeinen mit einer geringeren diskriminativen Kraft zu rechnen. Ein guter Kompromiss könnte eine Kombination aus beiden Optionen sein. Dabei würden zuerst die Atome, welche Schauplatz einer tautomeren oder protomeren Veränderung sein könnten (protomere bzw. tautomere Zentren), ermittelt und deren Atomtyp würde anschließend auf einen neuen speziell für protomere bzw. tautomere Zentren zugeschnittenen Atomtyp geändert. Damit könnte eine gleichbleibende Eingabegraph-Anzahl erzielt werden. Für die Bestimmung der protomeren bzw. tautomeren Zentren können neben der zeitintensiven Enumeration sicherlich ebenfalls günstigere Vorhersagealgorithmen genutzt werden.

Generell ist der Umgang mit der Stereoisomerie wichtig und nicht immer ausreichend bedacht. In Scaffold Hunter kann der Nutzer selbstständig festlegen, ob Stereoisomere als identisch angesehen werden sollen oder ob nach Stereoisomerie differenziert werden soll. Einhergehend mit der Stereoisomerie ist verallgemeinert die dreidimensionale Anordnung des Moleküls von Bedeutung für eine Interaktion mit einer biologischen Zielstruktur. Die dreidimensionale Anordnung wird in den in dieser Arbeit vorgestellten Graph-Repräsentationen nicht berücksichtigt. Ein Grund für diese Entscheidung war, wie bereits in der Shape-Ebene erwähnt, dass die Vorhersage der biologisch aktiven Konformation ein aktuell offenes Problem darstellt, welches zum jetzigen Zeitpunkt nicht effizient für eine große Anzahl an Molekülen und unabhängig von einer potentiellen Zielstruktur gelöst werden kann. Dennoch könnten in zukünftigen Arbeiten ebenfalls die dreidimensionalen Informationen in der Graph-Repräsentation des Moleküls abgebildet werden.



(a) Sildenafil (b) Sildenafil

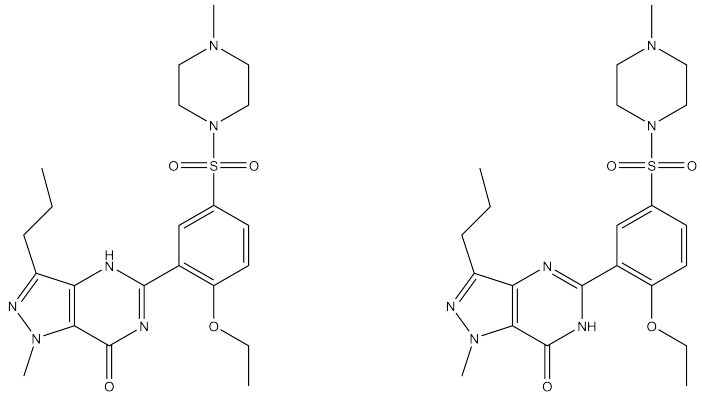
(c) Vardenafil (d) Vardenafil

Abbildung III.4.12.: Beispiele für Protomere anhand von Sildenafil (a) und (b) sowie Vardenafil (c) und (d).

In einer ersten Vorarbeit dazu von Gehde (2018) konnte eine mögliche Umsetzung für die Pharmakophor-Ebene gezeigt werden, da dort die dreidimensionale Anordnung von besonderer Bedeutung ist.

Das Ziel einer Ähnlichkeitsanalyse sollte zuvor definiert werden. Sollen ähnliche Moleküle unabhängig von einem vorgegebenen Targetprotein gefunden bzw. geclustert werden, so empfiehlt es sich, möglichst diverse Ansätze zur Bestimmung der Ähnlichkeit zu verwenden. Ist die Ähnlichkeit zu einem vordefinierten Targetprotein von Interesse, so lohnt sich die Optimierung der Ähnlichkeitsanalyse.

Zuletzt sei noch erwähnt, dass eine ausreichende Anzahl an qualitativ hochwertigen Bioaktivitätsdaten die fundamentale Grundlage für jedwede Bioähnlichkeitsanalyse ist. Insbesondere Daten zu inaktiven Verbindungen sind aktuell bedauerlicherweise noch unzureichend vorhanden.



(a) Sildenafil (b) Sildenafil

(c) Vardenafil (d) Vardenafil

Abbildung III.4.13.: Beispiele für Tautomere anhand von Sildenafil (b) und (a) sowie Vardenafil (d) und (c).

III.4.3. Zusammenfassung

Für die Bioisoster-Ebene wurde direkt die Bioähnlichkeit betrachtet. Es konnten 93.282 verlässliche bioisostere Paare identifiziert werden. Einige der identifizierten bioisosteren Paare wurden anhand von Arzneistoffen verdeutlicht. Eine erste Vorimplementierung der Bioisoster-Ebene konnte durch Kevin Nikiel umgesetzt werden (Nikiel 2017). Der nächste Schritt wird die Integration dieser Vorimplementierung in den bestehenden BBPMCS-Suchalgorithmus von Andre Droschinsky sein. Die Shape-Ebene wurde einerseits über die Verwendung von NPRs und andererseits mithilfe von FastROCS und CSD-Fragmenten analysiert. Die NPR-basierte Shape-Analyse benötigt eine dreidimensionale Repräsentation der Moleküle und ermöglicht eine Charakterisierung des Shapes eines Moleküls unabhängig von einer Referenzstruktur. Der Shape-Raum von Produkten zweier DEL-Reaktionen wurde so analysiert. Mithilfe von FastROCS wurde jeweils eine Distanzmatrix für CSD-Fragmente mit und ohne *Substituenten-Dummies* erstellt.

Bei der Implementierung der Bioisoster- und Shape-Ebene kamen einige Fragen und Herausforderungen auf. Konkret sind dies: (i) Isomerie, (ii) verfügbare Bioaktivitätsdaten, (iii) Generierung bzw. Wahl von dreidimensionalen Strukturen, (iv) Eingrenzung des Substitutionsraums und (v) Alignment-Probleme. Für diese Herausforderungen wurden Lösungsansätze erarbeitet und diskutiert.

In Zukunft sollten die Herausforderungen vermehrt Beachtung bei der Entwicklung und Evaluierung von Algorithmen zur Ähnlichkeitsanalyse finden und die Lösungsansätze sollten konsequent umgesetzt, erweitert und verbessert werden.

Quintessenz:

- 93.282 verlässliche bioisostere Paare wurden identifiziert
- zwei Shape-basierte Distanzmatrizen mit 15.615^2 (ohne *Substituenten-Dummies*) und 39.414^2 (mit *Substituenten-Dummies*) Einträgen wurden berechnet
- die NPR-basierte Analyse zeigt, dass zur Erstellung einer möglichst diversen Bibliothek diverse Edukte und Reaktionen verwendet werden sollten
- Herausforderungen: Isomerie, verfügbare Bioaktivitätsdaten, Generierung bzw. Wahl von dreidimensionalen Strukturen, Eingrenzung des Substitutionsraums und Alignment-Probleme

Teil IV.

Material und Methoden

IV.1. Verbrauchsmittel und Chemikalien

IV.1.1. Allgemeine Anmerkungen

Kommerziell erhältliche Chemikalien wurden ohne vorherige Aufreinigung verwendet. Die Reinheit und Identität der Testverbindungen wurde mittels *high performance liquid chromatography* (Hochleistungsflüssigkeitschromatographie) (HPLC) und UV-Vis- (Reinheit) oder MS-Detektor (Identität) überprüft. Sofern nicht anders angegeben wurden Chemikalien von Sigma Aldrich (St. Louis, USA), Carl Roth (Karlsruhe, DE), Alfa Aesar (Ward Hill, USA) sowie Acros Organics (Teil von Thermo Fisher Scientific, Waltham, USA) und Verbrauchsmittel von Sarstedt (Nümbrecht, DE) bezogen.

IV.1.2. Verbrauchsmittel

Kristallisationsplatten	EasyXtal 15-Well Plates (Qiagen Sciences)
Assayplatten:	
Quellplatten	384-Well Low Dead Volume Microplate - 384LDV (Labcyte)
Intermediatplatten	ECHO Qualified 384-Well Polypropylene Microplate (Labcyte)
Zielplatten	Proxi Plate-384 Plus (Perkin Elmer)
Ultrafiltration	Vivaspin 20 und 500 (Sartorius)
Kassette für Dispenser	ThermoFisher Scientific
Säulen:	
Nickel-NTA	Ni-NTA Superflow Cartridge 5 mL (Qiagen)
Größenausschluss	HiLoad 6/60 Superdex 75 (GE Healthcare)
HPLC-Säule	VP 125/21 Nucleodur C18 (Macherey-Nagel)
Kapillaren (Standard)	NanoTemper Technologies

IV.1.3. Chemikalien, Proteine und Bakterienstämme

Assaykit	EPIgeneous Binding Domain Assay Kit A BRD4(1) (CisBio)
Sulfapyridin	TCI
NT-647-NHS	NanoTemper Technologies
Histon H4 Peptid	AnaSpec
Glycerol	AppliChem
deuteriertes DMSO	VWR Chemicals
TEV-Protease	Protean
flüssiger Stickstoff	Linde
3-Methylbutylnitrit	TCI
Escherichia coli BL21 (DE3)	Novagen (Merck)

IV.2. Geräte

Plattenlesegerät/UV-Vis Spektrometer	SpectraMax Paradigm (Molecular Devices)
Dispenser	Multidrop (Thermo Fisher Scientific)
akustischer Dispenser	ECHO 520 (Labcyte)
Vortex	Heidolph und Vortex Genie 2 (Scientific Industries)
Plattenzentrifuge	Centrifuge 5810 (eppendorf)
Tischzentrifuge	Labnet International Inc.
Zentrifuge	Centrifuge 5415R (eppendorf)
Zentrifuge	Sorvall Evolution RC (Thermo Fisher Scientific)
Rotor	SA-600 Sorvall (Thermo Fisher Scientific)
Rotationsverdampfer	Totavapor R-200 (BÜCHI)
Wasserbad	Heating Bath B-490 (BÜCHI)
Vakuumkontrolle	Vacuum Controller V-800 (BÜCHI)
Vakuumpumpe	Chemistry Hybrid Pump RC6 (vacuumbrand)
Gefriertrocknungsanlage	Alpha 2-4 LD plus (Chirst)

hochauflösendes Massenspektrometer	LTQ-FT Orbitrap - ESI (Thermo Electron, jetzt Thermo Fisher Scientific)
präparative HPLC	1260 Infinity (Agilent Technologies)
Mikroskop	CX41RF (Olympus)
Pipetten	Eppendorf
Spektrophotometer	NanoDrop 2000 (ThermoFisher Scientific)
Fast protein liquid chromatography System	ÄKTApure (GE Healthcare Life Sciences Services)
Hochdruckhomogenisator	M-110L Microfluidizer Materials Processor (Microfluidics)
Inkubationsschüttler	Multitron Pro (Infors HT)
Eismaschine	AF100 (Scotsman)
Gelelektrophorese	PowerPac Universal (Bio-Rad)
Feinwaage	Cubis MSU 125P-100 DI (Sartorius)
Waage	CP225D (Sartorius) und PG5002-S Delta Range (Mettler Toledo)
Autoklav	Varioklav (Thermo Fisher Scientific)
MST-Gerät	Monolith NT.115 (NanoTemper Technologies)
qPCR	LightCycler 480 (Roche)
Rechner und Grafikkarten	
Arbeitsplatzrechner	Grafikkarten (NVIDIA): Tesla K20c und Quadro K4000, Prozessor (Intel): Xeon CPU E5-2680 v2 @ 2.80 GHz, RAM: 64 GB
Pipeline Pilot Server	Grafikkarten (NVIDIA): Tesla K20c und Quadro K2000, Prozessor (Intel): Xeon CPU E5-2690 0 @ 2.90 GHz, RAM: 160 GB
Cluster	8 Nodes mit 20 Cores Headnode: Prozessor (Intel): Xeon CPU E5-2609 v3 @ 1.90 GHz andere Nodes: Prozessor (Intel): Xeon CPU E5-2650 v3 @ 2.30 GHz

IV.1. Labor Experimente

IV.1.1. Proteinexpression von BRD4

Der von Prof. Stefan Knapp freundlicherweise zur Vefügung gestellte pNIC28-Bsa4 Vektor wurde in kompetente *Escherichia coli* (*E. coli*) BL21 (DE3) Zellen transformiert. Die heterologe Proteinexpression von BRD4(1) erfolgte anschließend in Luria-Bertani Medium (LB-Medium) unter Zusatz von 50 μg/mL Kanamycin und nach Induktion mit 0,1 mM Isopropyl-β-D-thiogalactopyranosid (IPTG) bei 18 ℃ wie in Filippakopoulos *et al.* (2012) und Filippakopoulos *et al.* (2010) beschrieben. Die Zellen wurden durch Zentrifugation für 15 min bei 4 ℃ und 3.000 rpm geerntet, in Lysis-Puffer (siehe Abschnitt B.8.1) resuspendiert und bei -21 ℃ eingefroren.

IV.1.2. Proteinaufreinigung von BRD4

Die in Lysis-Puffer eingefrorenen Zellen wurden auf Eis aufgetaut, dispergiert und mithilfe eines Hochdruckhomogenisators (mindestens drei Zyklen) lysiert. Das Lysat wurde bei 4 ℃ und 20.000 rpm für 30 min zentrifugiert. Der Überstand wurde anschließend zur weiteren Verwendung mit einer Spritze abgenommen und mit einem 0,2 μm Filter filtriert. Danach erfolgte die Proteinaufreinigung von BRD4(1) per *fast protein liquid chromatography* (FPLC) über eine Nickel-NTA-Affinitätschromatographie (Gradient 50 % 1 M Imidazol pH 8,0 in 20 min, Flussgeschwindigkeit 1 mL/min) und eine anschließende Größenausschlusschromatographie nach Filippakopoulos *et al.* (2012). Die Chromatographiesäulen wurden zuvor entsprechend der Säule entweder mit Lysis-Puffer (siehe Abschnitt B.8.1, Nickel-NTA-Chromatographie) oder mit Größenausschluss-Puffer (siehe Abschnitt B.8.2, Größenausschluss-Chromatographie) äquilibriert. Die gesammelten Fraktionen der Nickel-NTA-Chromatographie wurden vereinigt und der Polyhistidin-Tag (His-Tag) mittels TEV-Protease (1 mg TEV-Protease auf 100 mg BRD4) über Nacht bei 4 ℃ abgespalten. In der anschließenden Größenausschlusschromatographie (Flussgeschwindigkeit 1 mL/min) wurde der abgespaltene His-Tag aus der Proteinlösung entfernt. Dazu wurde die Proteinlösung zuvor auf ca. 2 mL aufkonzentriert. Die gesammelten Fraktionen der Größenausschlusschromatographie wurden vereinigt und auf 10-20 mg/mL im Größenausschluss-Puffer (siehe Abschnitt B.8.2) mittels Ultrafiltration aufkonzentriert. Falls nicht direkt mit dem gewonnenen Konzentrat weiter gearbeitet wurde, erfolgte die Lagerung bei -80 ℃. Die Proteinsequenzen mit und ohne His-Tag befinden sich im Anhang unter B.7.

IV.1.3. Proteinkristallographie von BRD4

Zur Verifizierung der gefundenen BRD4-Inhibitoren und zur Aufklärung des Bindemodus wurden Kokristallisationsexperimente durchgeführt. Dazu wurde BRD4(1) sofort nach der Aufreinigung (siehe Abschnitt IV.1.2) verwendet. Alle Kokristalle wuchsen innerhalb von zwei Wochen bei 4 ℃ bis zur Diffraktionsqualität. Für den Inhibitor **3** konnte ein Kokristall mittels der Dampfdiffusionsmethode im sitzenden Tropfen erhalten werden. Dazu wurden 100 nL einer BRD4(1)-**3**-Mischung (9 mg/mL BRD4(1) und 1 mM **3**), welche über Nacht bei 4 ℃ inkubiert wurde, mit 100 nL der Reservoirlösung-3 (siehe Anhang Abschnitt B.8.4) gemischt. Für die Inhibitoren **2**, **4**, Sulfasalazin (**22**) und SSLH01 (**21**) wurden die Kokristalle im hängenden Tropen durch die Dampfdiffusionsmethode erhalten. Es erfolgte eine zweistündige Inkubation der Protein-Inhibitor-Mischung bei 4 ℃. Die Protein-Inhibitor-Mischung bestand jeweils aus einer finalen BRD4(1)-Konzentration von 9 mg/mL und einer Inhhibitor-Konzentration von 5 mM für die Verbindungen **2** und **4** oder von 1 mM für die Verbindungen **21** und **22**. Die Protein-Inhibitor-Mischung wurde anschließend zur Abtrennung von ausgefallenem Protein oder Inhibitor zentrifugiert (10 min bei 13.000 rpm) und der Überstand weiter verwendet. Aus dem Überstand der Protein-Inhibitor-Mischung wurde 1 μL abgenommen und mit einem gleichen Volumen an Reservoirlösung vereinigt. Die Zusammensetzungen der verwendeten Reservoirlösungen befinden sich im Anhang unter B.8.4. Die gefischten Kristalle wurden in eine Kryoschutzlösung (Reservoirlösung versetzt mit 10 % Glycerol (Inhibitor **3**), 10 % Ethylenglycol (Inhibitoren **2** und **4**) oder 25 % Ethylenglycol (Inhibitoren **21** und **22**) überführt und für wenige Sekunden inkubiert. Anschließend wurden die Kristalle sofort in flüssigem Stickstoff eingefroren.

IV.1.4. Röntgenstrukturanalyse

Die Diffraktionsdaten der Kokristalle wurden bei 100 K mit der X10SA Röntgenquelle am Synchroton der Swiss Light Source (PSI, Villigen, Schweiz) aufgenommen. Die Röntgenstrahlwellenlänge betrug für alle Komplexstrukturen außer für die Kokristalle mit Inhibitor **4** und **21** ca. 1 Å. Zur Verifizierung der Lokalisation des Inhibitors **4** wurde das anomale Signal des Brom-Atoms bei der Peakwellenlänge (0,919740 Å) der Brom K-Kante aufgenommen. Für die Komplexstruktur mit Inhibitor **21** wurde eine Wellenlänge von 1,7699 Å verwendet, um das anomale Schwefelsignal zu messen, da der Inhibitor eine Sulfonsäureamid-Gruppe enthält. Der Datensatz wurde zur Verfeinerung des finalen Strukturmodells als nativer Datensatz, d. h. ohne Verwendung der anomalen Signale, prozessiert. Die Datensätze wurden mit XDS (Kabsch 2010) und AIMLESS aus der CCP4 Softwaresuite 7.0 (Winn *et al.* 2011) prozessiert. Für die Komplexstruktur mit Inhibitor **4** wurde neben XDS auch XDS app zur Datenprozessierung verwendet. Die Kokristallstrukturen wurden durch molekularen Ersatz gelöst.

Dazu wurde die publizierte BRD4(1)-JQ1-Komplexstruktur (PDB-ID: 3MXF, Filippakopoulos *et al.* (2010)) als Template für PhaserMR (CCP4 Softwaresuite 7.0, Winn *et al.* (2011)) verwendet. Die Koordinaten und geometrischen Einschränkungen (*restraints*) wurden mittels PHENIX.eLBOW (PHENIX 1.10.1) (Moriarty, Grosse-Kunstleve und Adams 2009) für die Inhibitoren **2**, **4**, **21** sowie **22** und mittels PRODRG-Server (Schüttelkopf und van Aalten 2004) für den Inhibitor **3** erstellt. Die Verfeinerung der Komplexstrukturen und die Modellbildung erfolgte mithilfe von PHENIX.refine (PHENIX 1.10.1) und COOT (Emsley und Cowtan 2004). Für die Komplexstrukturen mit den Inhibitoren **2** und **4** konnte die erfolgreiche Verfeinerung mittels PDB-REDO-Webserver (Joosten *et al.* 2014) verifiziert werden. Das SAD-Phasing für die Komplexstruktur mit dem Inhibitor **4** wurde mit den Programmen SOLVE (Terwilliger und Berendzen 1999) und PHASER (CCP4 Softwaresuite 7.0) durchgeführt. Die Komplexstrukturen von **21** und **22** wurden mithilfe von WhatCheck 10.1 (Vriend 1990) und MolProbity (Chen *et al.* 2010) validiert. Alle Strukturen wurden vor der Übersendung an die PDB mit dem PDB Validation Service (*wwPDB Validation Service* 2018) überprüft. Die Datensammlungs- und Verfeinerungsstatistiken befinden sich im Anhang unter B.6.1.

IV.1.5. Assaysysteme

Es wurden drei verschiedene Assaysysteme zur Ermittlung der inhibitorischen Wirkung der Analyten auf BRD4 verwendet. Der Primärscreen basierte auf einem HTRF-Assay. Die Ergebnisse wurden anschließend in zwei orthogonalen Assays (MST und TSA) validiert.

IV.1.5.1. BRD4 homogeneous time-resolved Förster-Resonanzenergietransfer (HTRF) Assay

Zur Bestimmung der Verdrängung des Histon H4 Peptides (kurz: H4 Peptid) aus der Bindung mit BRD4 durch eine zu testende Verbindung wurde ein Förster-Resonanzenergietransfer (FRET) basiertes Assaysystem gewählt. Dabei erfolgt eine Energieübertragung von einem angeregten Donor auf einen sich in der räumlichen Nähe befindenden Akzeptor. Der durch die emittierte Energie des Donors angeregte Akzeptor emittiert erneut Licht einer spezifischen Wellenlänge, welche mithilfe eines Detektors ausgewertet werden kann. Umgesetzt wurde dies durch das Epigeneous Binding Domain Kit (Cisbio), welches eine Spezialform des FRET nutzt. Von Vorteil ist, dass dieses Assaysystem für Hochdurchsatzscreenings geeignet ist und dass durch den Zeitversatz (*time-resolved*) zwischen Anregung und Fluoreszenzmessung unspezifische kurzlebige Fluoreszenzsignale (z. B. Eigenfluoreszenz des Analyten) seltener einen Einfluss auf das Messergebnis haben (*TR-FRET basics - Cisbio* 2018; *Epigeneous Bromodomain Assay: BRD4(1) - Cisbio* 2018).

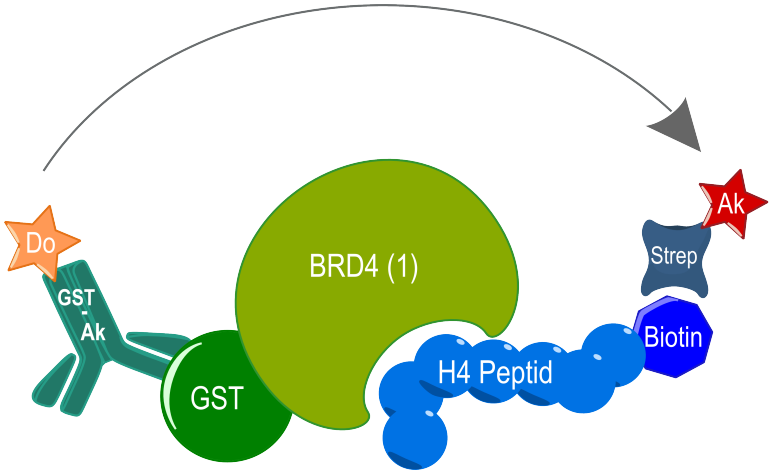


Abbildung IV.1.1.: HTRF-Assay Prinzip am Beispiel von BRD4. Ak: Akzeptor, Do: Donor, GST: Glutathion-S-Transferase, GST-Ak: Glutathion-S-Transferase Antikörper, Strep: Streptavidin.

Im Fall von BRD4 wird für das Assaysystem BRD4 mit einem GST-Tag (Glutathion-S-Transferase-Tag) und ein biotinyliertes Histon H4 Peptid verwendet. Der Energietransfer erfolgt über einen Europium-Kryptat-markierten Anti-GST-Antikörper (Donor) und Fluorophor-markiertes Streptavidin (Akzeptor) (siehe Abbildung IV.1.1). Erhöht sich der Abstand zwischen Donor und Akzeptor, das bedeutet zwischen BRD4 und H4 Peptid, so sinkt der Energietransfer und das detektierte Fluoreszenzsignal des Akzeptors. Dies kann bei der Bindung eines Analyten an BRD4, welche mit einer Verdrängung des H4 Peptides einhergeht, beobachtet werden. Berechnet wird das Verhältnis aus den Emissionssignalen von Akzeptor und Donor nach der Gleichung IV.1.1 (HTRF-Ratio).

$$HTRF - Ratio = \frac{10^4 \cdot \text{Akzeptorsignal bei 665 nm}}{\text{Donorsignal bei 620 nm}} \tag{IV.1.1}$$

Die Durchführung des Assays erfolgte in 384-Well ProxiPlates und mit einem finalen Assayvolumen von 15 μL unter Verwendung von Vormischungen (Premixe), wie in *Epigeneous Bromodomain Assay: BRD4(1) - Cisbio* (2018) beschrieben. Als Positivkontrolle dienten drei Substanzen, die *chemical probe* ± JQ1 (Filippakopoulos *et al.* 2010; Arrowsmith *et al.* 2015; *Chemical Probes* 2018) sowie zwei weitere bekannte BRD4-Inhibitoren **24** und **25** (Strukturen in Abbildung B.8). DMSO wurde als Negativkontrolle verwendet. Bedacht werden muss, dass DMSO selbst ein BRD4-Ligand ist (Philpott *et al.* 2011). Allerdings wurden alle Verbindungen in DMSO gelöst, weshalb durch die Verwendung von DMSO als Negativkontrolle der Effekt des Lösungsmittels auf die Ergebnisse mit berücksichtigt wurde. Die Verbindungen wurden mithilfe des ECHO 520 auf die Assayplatten transferiert und ggf. wurden zuvor Verdünnungen hergestellt. Die Premixe wurden entweder per Pipette oder durch Multidrop Dispenser auf die Assayplatten übertragen.

Ausgelesen wurden die Platten mithilfe eines SpectraMax Paradigm Plattenlesegeräts. Die IC_{50}-Werte wurden mithilfe der Software Quattro Workflow (quattro research) ermittelt. Zeigte eine Verbindung eine konzentrationsabhängige Veränderung in der Intensität der Fluoreszenz bei der Wellenlänge der Donoremission (Donorkanal), so wurde sie von weiteren Untersuchungen ausgeschlossen und als „nicht bestätigt“ gekennzeichnet.

Im initialen Screen wurde ein Hit definiert als eine Verbindung, welche bei einer Konzentration von 10 μM die verbleibende Bindungsaktiviät (kurz Restaktivität) auf 70 % oder 50 % reduziert. Als Qualitätskriterium wurde der Z'-Faktor verwendet. Bei einem Z'-Faktor von mindestens 0,5 kann von einem exzellenten Assay gesprochen werden, welcher für Hochdurchsatzsreenings geeignet ist. Der Z'-Faktor war mindestens 0,9 für jede Assayplatte. In einem anschließenden zweiten Experiment wurden von Hits aus dem initialen Screen IC_{50}-Werte ermittelt. Dazu wurde das HTRF-Ratio bei acht verschiedenen Konzentrationen zwischen 3 nM und 10 μM bestimmt. Zudem konnten somit ebenfalls Verbindungen mit einem kritischen Verhalten, d. h. konzentrationsabhängige Veränderungen im Donorkanal, identifiziert werden. Der Z'-Faktor lag bei mindestens 0,85 für jede Assayplatte.

Es folgten Screens zur Messung von IC_{50}-Werten als Triplikat und zur Bestimmung von IC_{50}-Werten mit Substanzkonzentrationen bis 50 μM (3 nM – 50 μM). Außerdem wurden manche Verbindungen nur bei den drei Konzentrationen: 10 μM, 30 μM und 50 μM gemessen. Der Z'-Faktor war in allen drei Experimentarten mindestens 0,8.

IV.1.5.2. BRD4 Microscale Thermophoresis (MST) Assay

Die Bindungsaffinitäten der BRD4-Inhibitoren wurden per MST durch Dr. Jette Pretzel bestimmt. Bei der MST wird ausgenutzt, dass sich Moleküle gerichtet entlang eines Temperaturgradienten bewegen (*Thermophorese*). Dadurch kommt es in den Bereichen mit einer erhöhten Temperatur zu einer geringeren Molekülkonzentration. Das Thermophoreseverhalten wird durch Veränderungen der Hydrathülle des Moleküls z. B. durch Ladungs- oder Molekülgrößen-Veränderungen, wie sie oft im Zusammenhang mit einem Bindungsvorgang beobachtet werden können, beeinflusst. Durch die Verwendung von Fluorophor-markierten Molekülen wird die Veränderung des Thermophoreseverhaltens insbesondere in Abhängigkeit von der Konzentration des Bindungspartners detektiert. Dazu wird die Veränderung der Fluoreszenzverteilung in mehreren Kapillaren mit jeweils einer Probe der Analyten-Verdünnungsreihe gemessen, nachdem eine lokale Temperaturerhöhung mittels Infrarot-Laser hervorgerufen wurde. Eine Beschreibung des MST-Assay Prinzips inklusive einiger Anwendungsbeispiele befindet sich in Jerabek-Willemsen *et al.* (2014).

Zur Bestimmung der Bindungsaffinitäten von BRD4-Inhibitoren wurden die primären Amine von aufgereinigtem BRD4-Protein (siehe Abschnitt IV.1.2) mit dem Fluoreszenzfarbstoff NT-647-NHS markiert. Die Konzentration von BRD4 wurde mit 200 nM konstant gehalten, wohingegen die Analyt-Konzentration von 3,1 nM bis 100 μM variiert wurde (insgesamt 16 Verdünnungen).

Die Verdünnungen erfolgten aus einem 10 mM DMSO-Stock in mit Polysorbat 20 versetztem (final 0,01 %) Größenausschluss-Puffer (Zusammensetzung siehe Anhang unter B.8.2). Die Inkubation von Protein und Analyt erfolgte im Dunkeln für 5 min bei 18 ℃. Anschließend wurde die Messung in Standardkapillaren bei 10 % LED-Stärke, 20 % MST-Stärke und 25 ℃ mittels Monolith NT.115 durchgeführt. Gemessen wird die normalisierte Fluoreszenz. Sie setzt sich einerseits aus der Temperaturabhängigkeit der Fluoreszenz des Farbstoffs und andererseits aus dem Konzentrationsverhältnis des Fluoreszenzfarbstoffs zwischen verschieden warmen Bereichen zusammen. Des Weiteren hat sowohl das ungebundene als auch das Inhibitor-gebundene Fluoreszenzfarbstoff-Protein-Konjugat einen Anteil an der normalisierten Fluoreszenz. Folglich kann der Anteil an Inhibitor-gebundenem Fluoreszenzfarbstoff-Protein-Konjugat über eine Verdünnungsreihe mithilfe der normalisierten Fluoreszenz ermittelt werden. Dazu wird die normalisierte Fluoreszenz gegen die Analyt-Konzentration aufgetragen. Mithilfe der K_d-Fitting-Funktion der NT Analyse Software des Monolith NT.115 wurden die so erhaltenen Kurven nach dem Massenwirkungsgesetz gefittet, um den K_d-Wert zu berechnen.

IV.1.5.3. BRD4 Thermal Shift Assay (TSA)

Mit steigender Temperatur erhöht sich in der Regel der Anteil an ungefaltetem Protein. Dadurch werden zunehmend hydrophobe Bereiche solvensexponiert. Durch die Bindung mit einem anderen Molekül, z. B. einem Inhibitor, kann sich die Hitzestabilität des Proteins erhöhen und eine Denaturierung erfolgt erst bei höheren Temperaturen. Dieser Vorgang kann somit zur Identifizierung von niedermolekularen Verbindungen, welche durch Bindung an das Protein dessen Hitzestabilität beeinflussen, genutzt werden. Als Fluorophor wird SYPRO-Orange verwendet, da zum einen die Extinktions- und Emissionswellenlängen mit dem PCR-Gerät kompatibel sind und es zum anderen unspezifisch an hydrophobe Bereiche bindet. Mit zunehmender Destabilisierung und Entfaltung des Proteins durch die Temperaturerhöhung nimmt die Fluoreszenz zu, bis sie durch die Aggregation des ungefalteten, denaturierten Proteins wieder sinkt. Die Messung der Inhibitor-induzierten Temperaturshifts von BRD4 erfolgte von Dr. Jette Pretzel nach Filippakopoulos *et al.* (2010). Dazu wurde aufgereinigtes BRD4-Protein (siehe Abschnitt IV.1.2) in TSA-Puffer (siehe Abschnitt B.8.3) mit einer Konzentration von 2 μM eingesetzt. Das finale Volumen eines Wells der 96-Well-PCR-Platte betrug 20 μL. Die Analyten wurden in einer Konzentration von 10 μM untersucht und SYPRO-Orange (5000-facher DMSO-Stock) wurde in einer Verdünnung von 1 zu 5000 verwendet. Die vorbereiteten Platten wurden mit Abdeckfolie versehen und für 2 min bei 500 g und 4 ℃ zentrifugiert. Die Temperaturerhöhung erfolgte von 25 ℃ auf 96 ℃ mit einem Gradienten von 3 ℃ pro Minute. Die Wellenlängenfilter wurden auf ca. 465 nm (Extinktion) und ca. 590 nm (Emission) eingestellt. Die Temperaturabhängigkeit der Fluoreszenzintensität von SYPRO-Orange wurde nach der Formel in Filippakopoulos *et al.* (2010) approximiert. Die Experimente wurden als vierfache Replikate durchgeführt und dreimal wiederholt.

IV.1.6. Synthese von 4-Amino-2-hydroxy-5-[4-(2-pyridylsulfamoyl)-phenyldiazenyl]-benzoesäure (SSLH01, 21)

Die Synthese von SSLH01 (**21**, Struktur siehe Abbildung IV.1.2) wurde nach Zhang *et al.* (2013) von Dr. Luis Bering und Saskia Spitzer durchgeführt. Anschließend wurde die methanolische Lösung filtriert und per präparativer Umkehrphasen HPLC (MeOH + 0,1 % TFA, Flussgeschwindigkeit 10 mL/min, Gradient 10 % MeOH auf 100 % MeOH in 30 min, Säule VP 125/21 Nucleodur C18 der Firma Macherey-Nagel) aufgereinigt. Die gesammelten Fraktionen wurden vereinigt und das organische Lösungsmittel im Rotationsverdampfer entfernt. Danach erfolgte die Lyophilisierung und die Lagerung als Lyophilisat bei 4 ℃ (zur Weiterverarbeitung) und danach bei -20 ℃ (Dauerlagerung). Die hochauflösende Masse von **21** wurde durch die AMD Einrichtung der Fakultät für Chemie und Chemische Biologie der TU Dortmund an einem Thermo Electron LTQ-FT Orbitrap (ESI) Massenspektrometer bestimmt. Für die Datenaufnahme und Datenverarbeitung wurde die Xcalibur 2.2 Software (Thermo Fisher) verwendet.

HRMS (ESI-MS) Berechnet: 414,08667 Da für $C18H15N5O5S^+$ $[M+H]^+$ m/z. Gefunden: 414.08638 Da

Zur weiteren Identitätsüberprüfung wurden NMR-Messungen an einem Bruker AV600 NMR-Spektrometer durch die NMR-Abteilung der TU Dortmund durchgeführt. Die Auswertung erfolgte durch Dr. Mauro S. Nogueira mithilfe von MestReNova Version 12.0.1-20560 (Mestrelab Research S.L.).

^{1}**H NMR** (600,3 MHz, MeOH-d4) δ 8,37 (s, 1H), 8,08 (d, J = 8,6 Hz, 2H), 8,00 (br d, J = 6,3 Hz, 1H), 7,95 (d, J = 8,6 Hz, 2H), 7,78 (m, J = 8,3 Hz ,J = 6,3, J = 2.4 Hz, J = 1,7 Hz, 1H), 7,33 (d, J = 8,8 Hz, 1H), 6,93 (t, J = 6,5 Hz, 1H), 6,26 (s, 1H)

^{13}**C NMR** (150,9 MHz, MeOH-d4) δ 173,8, 166,7, 156,6, 155,2, 152,6, 143,6, 143,5, 142,6, 135,7, 132,8, 129,5, 123,3, 117,1, 116,6, 104,7, 101,8.

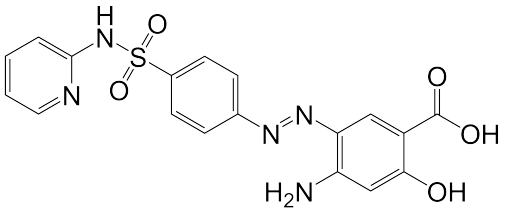


Abbildung IV.1.2.: Struktur von SSLH01 (**21**).

IV.2. *In silico* Experimente

Sofern nicht anders vermerkt, wurden die Standardeinstellungen verwendet. Alle Berechnungen mit dem Programm Pipeline Pilot (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015) wurden auf dem Pipeline Pilot Server durchgeführt. Alle anderen Berechnungen wurden, sofern nicht anders angegeben, auf dem Arbeitsplatzrechner durchgeführt.

IV.2.1. Erstellung von Plots und Histogrammen

Die Plots und Histogramme wurden mithilfe von Microsoft Office Professional Plus Excel, LibreOffice Calc, Gnuplot oder der Python-Bibliothek *Matplotlib* erstellt.

IV.2.2. Erstellung von Proteinabbildungen

Die Proteinabbildungen wurden mittels UCSF Chimera (Pettersen *et al.* 2004) generiert.

IV.2.3. Erstellung von Abbildungen von Reaktionen und niedermolekularen Verbindungen

Reaktionen wurden in MarvinSketch (ChemAxon) oder ChemDraw (PerkinElmer) gezeichnet und als rxn-Files gespeichert. Niedermolekulare Verbindungen wurden mithilfe von ChemDraw (PerkinElmer) gezeichnet.

IV.2.4. Erstellung von Workflows

Zur Erstellung von Workflows wurden die Workflow-Umgebungen Pipeline Pilot (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015) und KNIME (Berthold *et al.* 2008) verwendet. Workflow-Umgebungen erleichtern den Umgang mit Daten und ihre anschließende Analyse. Verschiedene Funktionen, z. B. das Einlesen einer Datei, werden durch entsprechende Komponenten bzw. Knoten durchgeführt. Diese Komponenten bzw. Knoten können miteinander verknüpft werden, wodurch ein Workflow entsteht. Zum Beispiel können in einem ersten Schritt die Daten eingelesen werden, in einem zweiten Schritt erfolgt eine Filterung der Daten (z. B. Molekülmasse kleiner 500 Da) und die bereinigten Daten werden in einem dritten Schritt in einer neuen Datei gespeichert.

IV.2.5. Vorbereitung von Molekülen und Reaktionen zur Erstellung von CH*I*PMUNK

Die Eduktdatenbanken zur Erstellung von CH*I*PMUNK wurden aus der MolPort, der eMolecules und der ZINC (All Now Subset) (2015) mittels MOE (*Molecular Operating Environment (MOE), 2013.08* 2013) durch Protonierung, Bindungslängenkorrektur (*wash-function*) und Energieminimierung gewonnen. In Pipeline Pilot (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015) wurden die Moleküle auf Fehler (z. B. Valenzfehler) überprüft. Fehlerhafte sowie doppelte Moleküle wurden entfernt. Um mindestens einen Umsatz je Reaktion zu garantieren, wurde jedes Edukt in seiner minimalen Form der Eduktdatenbank hinzugefügt. Durchgeführt wurde die Reaktion mithilfe der Pipeline Pilot Komponente *enumerate reaction*. Zur besseren Handhabbarkeit wurde vor der Durchführung der Reaktion die Menge der Edukte so gefiltert, dass nur noch Edukte mit den für die Reaktion benötigten funktionellen Gruppen beibehalten wurden.

IV.2.6. Verwendete Reaktionen zur Erstellung von CH*I*PMUNK

Die Reaktionen der medchem Sub-Bibliothek umfassen durch medizinische Chemiker bevorzugte Reaktionen und entstammen der Arbeit von Hartenfeller *et al.* (2011). Die heterozyklen Sub-Bibliothek beruht auf Produkten aus Reaktionen, welche häufig in Arzneistoffen anzutreffende Ringe darstellen. Eine entsprechende Auflistung häufig in Arzneistoffen anzutreffender Ringe befindet sich in Taylor, MacCoss und Lawson (2014). Anschließend wurden Reaktionen zur Darstellung dieser Ringe in SciFinder (*SciFinder* 2018) gesucht. Die dritte Sub-Bibliothek, genannt MCR, setzt sich aus den Produkten von Multikomponentenreaktionen (MCR) zusammen. Die verwendeten MCRs basieren auf den Reaktionen unter *Anchor-Biased MCR Reactions* (2018) aus der Arbeit von Koes, Dömling und Camacho (2018).

IV.2.7. Berechnung der physikochemischen Eigenschaften von CH*I*PMUNK

Pipeline Pilot 9.5 (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015) wurde zur Kalkulation der Molekülmasse, des AlogP-Wertes, der PSA, des SA-Scores, der Anzahl von Wasserstoffbrückendonatoren, Wasserstoffbrückenakzeptoren, aromatischen Bindungen und Doppelbindungen sowie der Konformität mit der Lipinski-Regel verwendet. Die Berechnung des SA-Scores basiert auf der Arbeit von Ertl und Schuffenhauer (2009), allerdings resultiert die Implementierung in Pipeline Pilot in einer anderen Normierung.

Der SA-Score von Pipeline Pilot besitzt in der Regel Werte zwischen -1 und +1. Dabei weisen höhere Werte auf eine leichtere synthetische Zugänglichkeit der Moleküle hin (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015).

IV.2.8. Clustering

Das Ziel eines Clusterings ist es eine Menge in Submengen zu unterteilen, sodass sich die Daten (hier Moleküle) innerhalb einer Submenge mehr ähneln als Daten (Moleküle) aus verschiedenen Submengen. In dieser Arbeit wurde der von Till Schäfer entwickelte und implementierte StruClus-Algorithmus (siehe Abschnitt II.1.3 und Schäfer und Mutzel (2017)) als Clustering-Algorithmus verwendet.

IV.2.8.1. Qualitätskriterien

Zur Kontrolle der Qualität eines Clusterings können verschiedene Qualitätskriterien verwendet werden. Es wird zwischen internen und externen Qualitätskriterien unterschieden. Für externe Qualitätskriterien wird eine Vorlage (*ground truth*) benötigt, welche das erwartete Ergebnis des Clusterings repräsentiert. In dieser Arbeit wurden primär externe Qualitätskriterien verwendet, um z. B. zu überprüfen wie gut die Einteilung der Moleküle durch StruClus mit der Einteilung nach Zielstruktur übereinstimmt. Dazu wurden der F1-Score (siehe unten) und die *Reinheit* (siehe unten) verwendet. Allerdings ist anzumerken, dass es kein ideales Qualitätskriterium gibt und jedes nur einen Teilaspekt der Qualität widerspiegelt. Zudem werden die Qualitätskriterien häufig (wie auch in dieser Arbeit) gewichtet, um bestimmte Einflüsse auf das jeweilige Qualitätskriterium bezüglich der Gesamtqualität besser zu normalisieren.

F1-Score Der F1-Score ist ein externes Qualitätskriterium, welches die *ground truth* als Grundlage zur Bewertung der Güte des Clusterings verwendet. Einerseits wird berücksichtigt, wie hoch der Anteil in dem zu untersuchenden Cluster (C_i) an Molekülen ist, die aus demselben *ground truth* Cluster (C_{gt}) stammen (Precision, siehe Gleichung IV.2.1) und andererseits wie hoch der Anteil an Molekülen aus dem *ground truth* Cluster ist, die in dem zu untersuchenden Cluster wiedergefunden werden können (Recall, siehe Gleichung IV.2.2). Der Wertebereich liegt zwischen null und eins. Ein Wert von eins bedeutet eine mit der *ground truth* identische Einteilung, d. h. identische Cluster, durch den getesteten Clustering-Algorithmus.

$$Precision_{C_i} = \frac{C_i \cap C_{gt}}{C_i} \tag{IV.2.1}$$

$$Recall_{C_i} = \frac{C_i \cap C_{gt}}{C_{gt}} \tag{IV.2.2}$$

Der Gesamt-Recall und die Gesamt-Precision werden jeweils gewichtet und ungewichtet berechnet. In den Gleichungen IV.2.3 (ungewichtet) und IV.2.4 (gewichtet) ist die Berechnung beispielhaft für den Gesamt-Recall gegeben mit n = Anzahl der Cluster und m_i = Anzahl der Clusterangehörigen des Clusters i.

$$Recall_{gesamt,ungewichtet} = \sum_{i=1}^{n} Recall_i \cdot \frac{1}{n}, mit\ i \in Menge\ der\ Cluster \tag{IV.2.3}$$

$$Recall_{gesamt,gewichtet} = \sum_{i=1}^{n} Recall_i \cdot \frac{m_i}{\sum_{i=1}^{n} m_i}, mit\ i \in Menge\ der\ Cluster \tag{IV.2.4}$$

In die Gleichung zur Berechnung des F1-Scores (siehe Gleichung IV.2.5 und Larsen und Aone (1999)) geht dann entweder der ungewichtete Recall und die ungewichtete Precision (siehe Gleichung IV.2.3) ein oder der gewichtete Recall und die gewichtete Precision (siehe Gleichung IV.2.4). Im ersten Fall ergibt sich der ungewichtete F1-Score und im zweiten Fall resultiert ein gewichteter F1-Score.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{IV.2.5}$$

Reinheit Die Reinheit P_{xy} eines Clusters x gibt an, wie hoch der Anteil einer *ground truth* Klasse y an den Clusterangehörigen (m_x) ist. P_{xy} kann mit der Gleichung IV.2.6 berechnet werden.

$$p_{xy} = \frac{|m_{xy}|}{|m_x|},\ wobei\ m_{xy}\ =\ m_x\ \cap\ y \tag{IV.2.6}$$

Es ergibt sich ein Wertebereich von null bis eins, wobei höhere Werte ein im Bezug auf die Reinheit besseres Clusteringergebnis bedeuten. Zur Berechnung einer Gesamt-Reinheit über alle Cluster wird üblicherweise pro Cluster nur der maximale Reinheitswert verwendet und über alle Cluster aufsummiert sowie anschließend durch die Cluster-Anzahl n gemittelt (siehe Gleichung IV.2.7).

$$p_{gesamt} = \frac{\sum_{x=1}^{n} maxP_x}{n} \tag{IV.2.7}$$

In dieser Arbeit ist die Reinheit nach *ground truth* Klasse von Bedeutung, weshalb zur Berechnung statt der maximalen Reinheit die Reinheit separat für jede *ground truth* Klasse ermittelt wird. Die ungewichtete Gesamt-Reinheit für die *ground truth* Klasse y lässt sich somit durch

$$p_{gesamt,y} = \frac{\sum_{x=1}^{n} P_{xy}}{n_y} \tag{IV.2.8}$$

bestimmen. Dabei ist n_y die Anzahl der Cluster, die Objekte (hier Moleküle) der *ground truth* Klasse y beinhalten.

Die gewichtete Gesamt-Reinheit nach *ground truth* Klasse y setzt sich aus mehreren Gewichtungstermen zusammen. Zum einen wird der Anteil der Objekte (hier Moleküle) der *ground truth* Klasse y des Clusters x an allen Objekten der *ground truth* Klasse y ($K_{w,x,y}$) berücksichtigt.

$$K_{w,x,y} = (\frac{|m_{xy}|}{|y|})^2 \tag{IV.2.9}$$

Zum anderen werden die mittlere Clustergröße und eine Abschätzung der Clustergrößen-Verteilung beachtet. Die finale gewichtete Gesamt-Reinheit nach *ground truth* Klasse y ergibt sich durch:

$$Reinheit_{gesamt,gewichtet,y} = \sum_{x=1}^{n}(K_{w,x,y} \cdot P_{yx}) \cdot (\frac{min_y}{max_y \cdot n_y}). \tag{IV.2.10}$$

Wobei min_y bzw. max_y die minimale respektive maximale Clustergröße eines Clusters mit Objekten aus der *ground truth* Klasse y ist.

IV.2.8.2. StruClus Vortest und Test verschiedener Molekülrepräsentationen auf dem SBDS

Bevor StruClus auf CH*I*PMUNK angewendet wurde, erfolgte ein Vortest auf dem deutlich kleineren SBDS. Die Moleküle wurden mithilfe eines Python-Skriptes (siehe Abschnitt IV.2.10) in ein für StruClus kompatibles Format (gml) konvertiert und verschiedene Molekülrepräsentationen erstellt. Aufgrund der geringeren Datensatzgröße war auch ein Vergleich zum häufig in der Chemieinformatik verwendeten Jarvis-Patrick-Clustering möglich. Dazu wurden die Parameter *Threshold* (Schwellenwert) und *Rmin* (minimales gleiches Verhältnis) der ChemAxon-Implementierung des Jarvis-Patrick-Clusteringalgorithmus (*Jarvis-Patrick-Clustering Documentation ChemAxon* 2018) in 0,05er Schritten in Bezug auf den F1-Score optimiert. Aus der Optimierung resultierte ein Schwellenwert von 0,65 sowie ein minimales gleiches Verhältnis von 0,4. StruClus wurde hingegen mit den Einstellungen, wie in Schäfer und Mutzel (2017) beschrieben, verwendet, mit Ausnahme der minimalen Separation, für welche ein Wert von 0,2 verwendet wurde und der Größe der gemeinsamen Substruktur, welche auf 40 Knoten begrenzt wurde.

IV.2.8.3. StruClus-Clustering von CH*I*PMUNK

Jede Sub-Bibliothek von CH*I*PMUNK wurde zusammen mit den Edukt-Datenbanken (CoMol und ZINC) und ChEMBL Version 22.1 mittels des StruClus-Alorithmus geclustert. Damit ist sowohl der neue chemische Raum als auch die Überlappung mit dem bereits synthetisierten und/oder bioaktiven Raum sichtbar. Eine detaillierte Beschreibung des StruClus-Algorithmus findet sich in Schäfer und Mutzel (2017).

Zusammengefasst ist StruClus ein Struktur-basierter Clustering-Algorithmus für große Datensätze (Big Data) von kleinen gelabelten Graphen, wie z. B. Molekülgraphen. Dazu wird der Datensatz basierend auf gemeinsamen Substrukturen gruppiert (geclustert). Diese gemeinsamen Substrukturen werden zugleich als intuitive Beschreibung der Gruppe (des Clusters) verwendet. Des Weiteren wird sichergestellt, dass die Cluster gut voneinander separiert sind. Dies bedeutet, dass die Moleküle aus verschiedenen Clustern weniger ähnlich sind, als die Moleküle innerhalb eines Clusters. Für große Datensätze ist StruClus durch seine lineare Laufzeit in Bezug auf die Datensatzgröße, d. h. Anzahl der Moleküle, geeignet. Als Label wurden in den Versuchen mit StruClus, falls nicht anders angegeben, die Atomtypen oder Elemente für die Knoten und der Bindungstyp für die Kanten verwendet. Einen kleinen Überblick gibt zudem Abschnitt II.1.3.

IV.2.8.4. Visualisierung der Clustering-Ergebnisse

Zur Visualisierung des StruClus-Clustering-Ergebnisses von CH*I*PMUNK wurde der relative Anteil berechnet. Dieser kann interpretiert werden als der Faktor des Erwartungswertes, d. h. wie aussagekräftig die beobachtete Anzahl an Molekülen einer bestimmten Ursprungsdatenbank ist in Relation zum Erwartungswert bei einer zufälligen Selektion und berechnet sich nach der Gleichung IV.2.11.

$$H_{D,Cl} = \frac{|D \cap Cl|}{E(|D \cap Cl|)},\ mit\ D \in Datenbanken \tag{IV.2.11}$$

Wobei $Datenbanken = ZINC, CoMol, ChEMBL, CHIPMUNK - Sub - Bibliothek$ und Sub-Bibliothek entweder heterozyklen, medchem oder MCR meint. Cl ist die Menge der Cluster-Mitglieder und E der Erwartungswert. Final wird das Ergebnis der Gleichung IV.2.11, wie in Gleichung IV.2.12 gezeigt, normalisiert und das Resultat gegen die Cluster-ID in einem Plot aufgetragen.

$$score_{H_{D,Cl}} = \frac{H_{D,Cl}}{\sum_y H_{y,Cl}},\ mit\ y \in Datenbanken \tag{IV.2.12}$$

IV.2.9. KNIME-Workflow zur Suche nach Molekülen in CH*I*PMUNK

Das zu suchende Moleküle wird mithilfe des Knotens *MarvinSketch* gezeichnet. Die Datenbank wird über den Knoten *SDF Reader* eingelesen und anschließend in RDKit Objekte durch den *RDKit From Molecule* Knoten überführt. Danach erfolgt die Berechnung der Fingerprints für das Query-Molekül und die Datenbankmoleküle durch den Knoten *RDKit Fingerprint*. Die Fingerprints werden anschließend mithilfe des Knotens *Similarity Search* verglichen und ähnliche Moleküle ermittelt.

Abschließend werden die Daten aus der Datenbank mit den gefundenen Hits per *Joiner* Knoten zusammengeführt und im sdf-Format mithilfe des *SDF Writer* Knotens abgespeichert. Der Workflow ist in Abbildung IV.2.1 dargestellt.

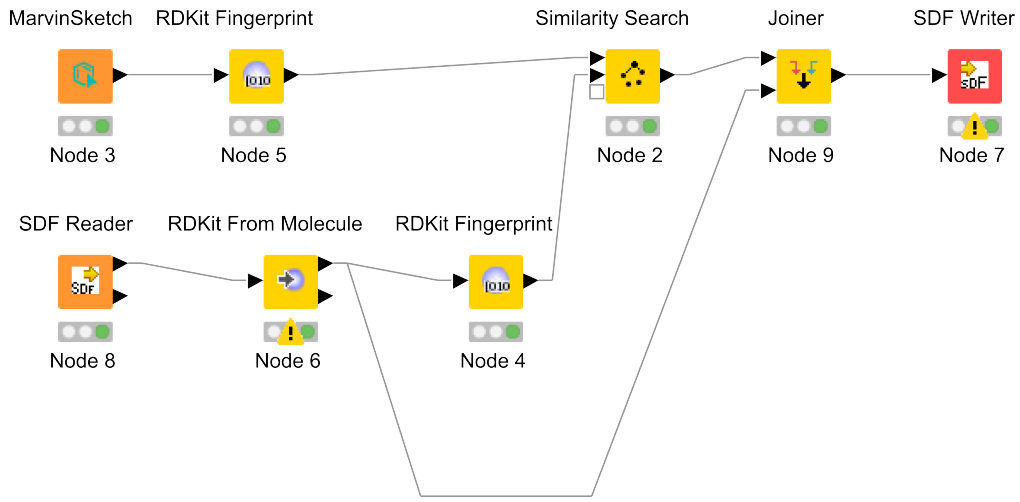


Abbildung IV.2.1.: Beispiel eines KNIME-Workflows zur Suche in CH*I*PMUNK.

IV.2.10. Vorbereitung der Moleküle für den BBPMCS-Algorithmus

Ein Python-Skript, welches auch die Atomtypengruppen zuordnen kann, wurde entwickelt, um Moleküle im mol2-Format in das für den BBPMCS-Algorithmus benötigte Eingabeformat gml zu konvertieren. Es erfolgte eine Weiterentwicklung dieses Python-Skriptes durch Robert Gehde, um benachbarte Knoten mit demselben Label zusammenzufassen (Gehde 2018). Die Zusammenfassung kann auf zwei Arten durchgeführt werden, entweder unter Beibehaltung der Wasserstoffatome oder ohne Speicherung dieser.

IV.2.11. Analyse des MSS-Ansatzes mithilfe des OBDS

Die durch O'Boyle und Sayle (2016) bereitgestellten Python-Skripte wurden so modifiziert, dass die Ergebnisse des BBPMCS-Algorithmus ebenfalls verarbeitet werden konnten. Darüber hinaus wurden die Skripte parallelisiert und so strukturiert, dass mehrere Versuche parallel getestet werden konnten. Die Berechnungen wurden auf dem Cluster durchgeführt.

IV.2.11.1. Änlichkeitssuche mittels ECFP4 und Tanimoto-Koeffizient (TC)

Die Änlichkeitssuche mittels ECFP4 und TC für die Proof of Concept Analyse unter Abschnitt III.2.2.1 wurde durch einen Pipeline Pilot 9.5 Workflow (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015) realisiert. Zuerst wurden die Moleküle der selektierten ChEMBL Version 22 (Bento *et al.* (2014), Selektion siehe Abschnitt IV.2.21) und Atenolol per *Mol2-Reader* eingelesen. Anschließend wurden doppelte Moleküle entfernt (Komponente *Remove Duplicate Molecules*) und Atenolol als Referenz definiert mithilfe der Komponente *Custom Manipulator(PilotScript)* sowie der Einstellung *Expression: IsReference:=true*. Die selektierten ChEMBL-Moleküle wurden ebenfalls mit der Komponente *Custom Manipulator(PilotScript)* verarbeitet und die Einstellung *Expression: true* hinzugefügt. Die Berechnung des ECFP4-TCs erfolgte durch die Komponente *Molecular Similarity (Tanimoto etc)* und mit den Einstellungen: (i) *SimilarityCoefficient - OutputContributions*: *True*, (ii) *UseProperties - PredefinedSet*: ECFP4, (iii) *MaximumNumberOfClosest*: 1, (iv) *MaximumNumberOfClosest - MinimumSimilarity*: 0, (v) *MaximumNumberOfClosest - MaximumSimilarity*: 1, (vi) *TestForReference: IsReference is defined AND IsReference*, (vii) *OptimizeFor*: *Speed* und (viii) *OptimizeFor - BitsetSize*: 1048. Nachfolgend wurden zur Kontrolle der *Custom Filter (PilotScript)* mit der Einstellung *Expression: ClosestSimilarities is defined* und die Komponente *Property Value Treshold Filter* mit den Einstellungen: (i) *PropertyName*: *ClosestSimilarities*, (ii) Filter: >=, (iii) *Threshold*: 0.0 und (iv) *FilterAs*: String verwendet. Abschließend wurden die Ergebnisse mittels des *Isentris CSV Writers* unter dem bei *Destination* angegebenen Pfad abgespeichert.

IV.2.11.2. Änlichkeitssuche mittels BBPMCS-Algorithmus

Der von Andre Droschinsky entwickelte und implementierte BBPMCS-Algorithmus wurde mit verschiedenen Parametern verwendet (Droschinsky, Kriege und Mutzel 2017; Droschinsky, Kriege und Mutzel 2018). Für die Proof of Concept Studie (siehe Abschnitt III.2.2.1) wurde das BBPMCS-Programm mit den Parametern: (i) -i Pfad zu den Eingabedateien, zuerst Query (zu suchendes Molekül/Graph) und dann Datenbank in der gesucht werden soll, (ii) -o Pfad zur Ausgabedatei und (iii) -t Anzahl der Berechnungsthreads aufgerufen. Die Benchmarkstudien auf dem OBDS erfolgten mit den Parameter: (i) -cto Ausgabe der Anzahl der Cliqueberechnungen, welche das Zeitlimit überschreiten in der Ausgabedatei, (ii) -ct Zeitlimit für Cliqueberechnungen in Millisekunden (getestet wurden 1000 ms und 2000 ms), (iii) -oneToX vergleicht jedes x+1te Molekül zu den x dazwischenliegenden Molekülen (verwendet wurde 4) und (iv) -o Pfad zur Ausgabedatei. Außerdem wurden in einigen Experimenten weitere Parameter wie: (i) -sc zu verwendende Ähnlichkeitsfunktion, (ii) -p Gap-Strafe und (iii) -l Pfad zur Ähnlichkeitsmatrix verwendet und entsprechend bei den Ergebnissen angegeben.

IV.2.12. Bindetaschenvergleich der APSC-Proteine

Der Bindetaschenvergleich wurde mithilfe von IsoMIF (Chartier und Najmanovich 2015; Chartier, Adriansen und Najmanovich 2016) von Christiane Ehrt, wie in Ehrt, Brinkjost und Koch (2018) beschrieben, durchgeführt. Dazu wurde ein 1082 Strukturen umfassendes Subset der sc-PDB (Kellenberger *et al.* 2006; Ehrt, Brinkjost und Koch 2018) verwendet und die Proteinstrukturen der APSC-Proteine hinzugefügt (PDB-IDs: 4O71 (BRD4), 3SZ1 (PPARγ), 2DUV (CDK2), 3DCU (NR1H4), 1X7J (ESR2)). Als Referenzstruktur gegen die verglichen wurde, wurde 4O71 verwendet. Zudem wurden Komplexstrukturen mit Luteolin (PDB-IDs: 4DGN, 5AUU, 5II2, 4QXV, 4HKN, 5NDF), Kinasestrukturen (PDB-IDs: 1CX4, 4DGN, 5AUU), zwei weitere Bromodomain-containing-Protein-Komplexstrukturen (PDB-IDs: 5II2, 3U5L) sowie eine Komplexstruktur mit Flavopiridol (PDB-ID: 3EBP) hinzugefügt. Es wurden zwei Ansätze durchgeführt. Einerseits wurden die vollständigen Liganden der APSC-Proteine verwendet und andererseits wurde nur das Scaffold (2- bzw. 3-Phenyl-4H-chromen oder 2-Phenylnaphthalin) berücksichtigt. Dies hat einen Einfluss auf die Bindetaschendefinition, welche IsoMIF zur Berechnung der Interaktionsfelder verwendet.

IV.2.13. Docking

Das Docking ist eine computer-basierte Methode zur Vorhersage eines potentiellen Bindemodus einer niedermolekularen Verbindung in einer Proteinbindetasche. Dabei wird eine sogenannte Pose der niedermolekularen Verbindung in der Proteinbindetasche generiert. Anschließend erfolgt die Bewertung der generierten Pose mittels einer Scoring-Funktion (für Details siehe Kitchen *et al.* (2004)). Die Dockingstudien in dieser Arbeit wurden mit der Software GOLD 5.2 (Jones *et al.* 1997) durchgeführt. Der Dockingansatz zur Identifizierung neuer BRD4-Inhibitoren wurde dreifach auf einem Cluster unter Verwendung der Scoringfunktion ChemScore (Eldridge *et al.* 1997; Baxter *et al.* 1998) durchgeführt. Als Zielstrukturen des Dockings wurden alle fünf mittels APSC gefundenen Proteine (BRD4, ESR2, CDK2, PPARγ und NR1H4) verwendet (*Ensemble-Docking*). Für Verbindungen die einen höheren Docking-Score für BRD4 als für die anderen Proteine des APSC erzielen konnten, wird angenommen, dass sie mit höher Wahrscheinlichkeit BRD4-Inhibitoren sind. Dadurch erfolgt eine restriktivere Auswahl mit potentiell erhöhter Wahrscheinlichkeit einer besseren Hitrate im *Ensemble-Docking* im Vergleich zu einem Docking ausschließlich in BRD4. Zuerst wurde ein freies Docking, ohne Restriktionen (*constraints*) durchgefürt. Dabei wurden für BRD4 die Seitenketten der Aminosäuren Tyr97, Asn140 und Asp145 flexibel, d. h. die Konformation kann aus einer Rotamerbibliothek gewählt werden, gesetzt. Anschließend erfolgte eine erneute Bewertung der gefundenen Posen (*Rescorring*) unter Verwendung von Restriktionen. Die Restriktionen bezogen sich auf die Aminosäuren mit denen Wechselwirkungen aus der Literatur bekannt waren, dies sind Asn140, Tyr97 und Asp145 (Ember *et al.* 2014). Der finale Score wurde durch eine Normalisierung (Carta, Knox und Lloyd 2007) erhalten.

IV.2.13.1. Vorbereitung der Moleküle für das Docking

Zur Vorbereitung der Moleküle für das Docking wurde die Software MOE 2013 (*Molecular Operating Environment (MOE), 2013.08* 2013) verwendet. Die Moleküle wurden mithilfe der *wash-Funktion* und der Einstellung *scale to reasonable bond length* protoniert. Zu dem protonierten Molekül wurde anschließend per *Energy Minimize-Funktion* mit einem RMS-Gradienten von 0,05 eine Konformation berechnet. Dabei wurde durch die Auswahl der Option *preserve existing chirality* die Stereoinformation beibehalten. Nicht verwendet wurde die Option *Add Hydrogens to Structure*. Zur leichteren Zuordnung der Docking-Ergebnisse zu den Molekülen wurde ihre ID (COMAS-ID) gespeichert. Dazu wurde die Funktion *Molecule Name* mit der Einstellung *store in molecule* verwendet und unter *Field* die Bezeichnung der Spalte, in der die COMAS-ID gespeichert wurde, angegeben sowie die Optionen *Copy name to each chain name*, *tag* und *header* ausgewählt. Final wurden die so vorbereiteten Moleküle im mol2-Format gespeichert.

IV.2.13.2. Vorbereitung der Proteine für das Docking

Zur Vorbereitung der Proteine für das Docking wurde die Software MOE 2013 (*Molecular Operating Environment (MOE), 2013.08* 2013) verwendet. Zuerst wurde die am besten geeignete Kette (z. B. möglichst vollständig aufgelöst in der Kristallstruktur) je Protein ausgewählt. Nicht mehr benötigte Ketten wurden über die Sequenz-Ansicht gelöscht. Danach erfolgte eine Überprüfung und gegebenenfalls eine Korrektur der Proteinstruktur mithilfe der Funktionen *Structure Preparation* und *Correct*. Anschließend wurde durch die Funktion *Protonate3D* das Protein virtuell protoniert. Abschließend erfolgte eine manuelle Kontrolle der Struktur und gegebenenfalls eine Durchführung weiterer Bearbeitungsschritte sowie eine Speicherung der so vorbereiteten Proteine im pdb- oder mol2-Format.

IV.2.14. Selektion von zu APSC Liganden ähnlichen Verbindungen

Aus der ChEMBL 19 (Bento *et al.* 2014) wurden mittels eines Pipeline Pilot (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015) Workflows Liganden der Proteine des APSC (ESR2, CDK2, PPARγ, NR1H4) selektiert. Dabei wurde ein SQL-Befehl, welcher nach den UniProt-IDs (The UniProt Consortium 2017; *UniProt* 2018): (i) Q92731, (ii) P24941, (iii) P37231 und (iv) Q96RI1 filterte, ausgeführt. Zu den so gewonnenen APSC-Liganden aus der ChEMBL wurden durch eine Ähnlichkeitssuche anhand des ECFP4-TCs ähnliche Verbindungen in der COMAS-Bibliothek gesucht. COMAS-Verbindungen, welche mindestens einen TC von 0,7 zu mindestens einem der aus ChEMBL gewonnen APSC-Liganden besitzen, wurden selektiert. Dazu wurde erneut ein Pipeline Pilot Workflow entworfen (analog zu dem Workflow unter Abschnitt IV.2.11.1) und ausgeführt.

Das Ergebnis waren 1.158 COMAS-Verbindungen mit der geforderten Ähnlichkeit zu den APSC-Liganden. Diese 1.158 Verbindungen wurden in einem HTRF-Screen (siehe Abschnitt IV.1.5.1) bezüglich ihrer Wirkung auf BRD4 untersucht. Das Vorgehen ist in einem Schema in der Abbildung IV.2.2 veranschaulicht. Ebenfalls wird das APSC visualisiert, um so einen Transfer des abstrakten Konzeptes (oberere Grafik) in die praktische Umsetzung und Validierung (untere Grafik) zu erleichtern.

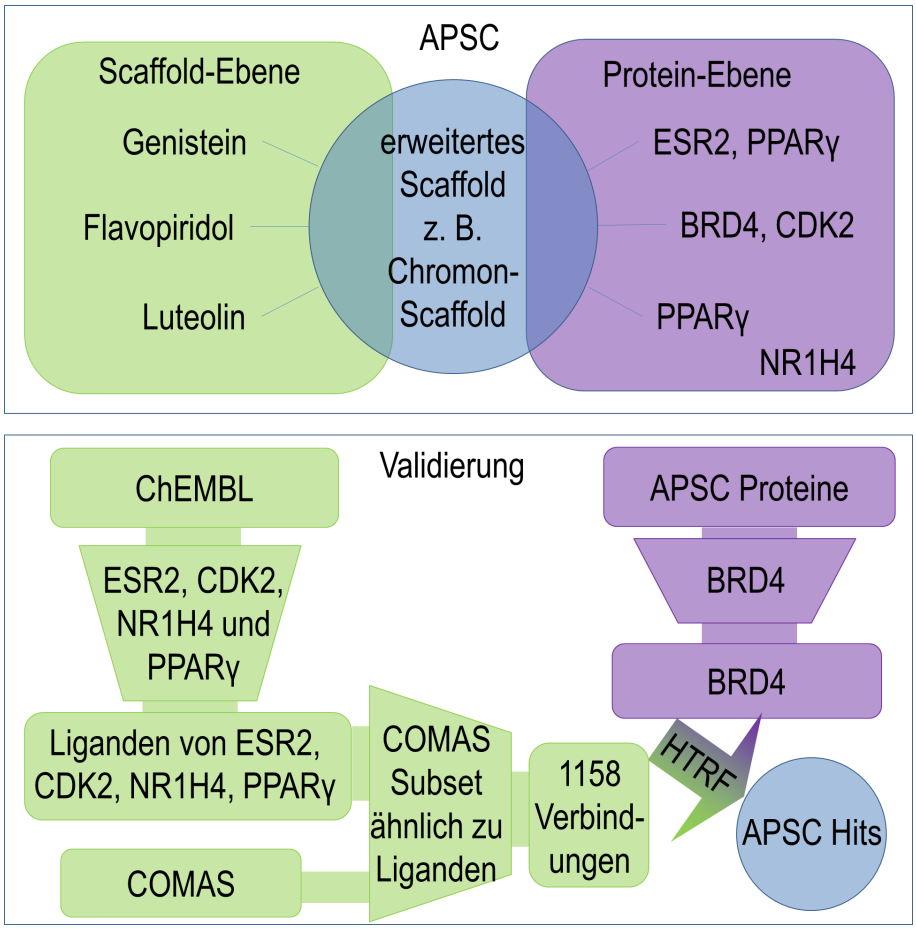


Abbildung IV.2.2.: Das APSC und seine Validierung.

IV.2.15. Suche nach Liganden von PPARγ oder BRD4

Zum Vergleich der gefundenen BRD4-Hits aus dem APSC und dem Docking-Ansatz mit bekannten Liganden für BRD4 und PPARγ wurden zuerst die BRD4- und PPARγ-Liganden aus der ChEMBL (Version 22) per KNIME-Workflow (Berthold *et al.* 2008) herausgefiltert. Dazu wurde der Knoten *ChEMBLdb Connector* mit den Optionen *Protein Accession* = O60885 bzw. P37231 (UniProt-IDs für BRD4 bzw. PPARγ) und *Search Type* = *target* sowie ohne Eintrag unter der Option *ChEMBL ID* verwendet und das Kontrollkästchen *Bioactivities* aktiviert. Anschließend wurden die gefundenen Substanz-ChEMBL-IDs mit dem Knoten *ChEMBLdb Connector Input* verarbeitet (u. a. wurde der SMILES abgerufen). Dabei wurden die Optionen *Search Type* = *compound* und *ChEMBL ID* unter Angabe der Spalte, in der die Substanz-ChEMBL-IDs gespeichert wurden, verwendet. Das Ergebnis wurde als csv-Datei gespeichert.

IV.2.16. Ermittlung der bezüglich EGFR-BRD4-Inhibition sensitiven und insensitven Zelllinien

Zur Ermittlung der Zelllinien, welche auf die duale EGFR-BRD4-Inhibition sensitiver reagieren als auf EGFR-Inhibition, wurde zuerst für jede Zelllinie die mittlere Proliferation ausgerechnet. Dazu wurden drei verschiedene mittlere Proliferationswerte berechnet. Erstens wurden nur die Proliferationswerte nach Behandlung mit EGFR-Inhibitoren (ohne RL078, RL127 und RL129) gemittelt ($\emptyset G_{EGFR}$). Zweitens wurden alle Proliferationswerte nach Behandlung mit einem Inhibitor der EGFR-Inhibitor-Serie (inklusive der EGFR-BRD4-Inhibitoren) verwendet ($\emptyset G_{gesamt}$) und drittens wurden die Proliferationswerte nach Behandlung mit EGFR-BRD4-Inhibitoren (RL078 und RL129) gemittelt ($\emptyset G_{dual}$). Entsprechend wurde mit der Standardabweichung verfahren (σG_{EGFR}, σG_{gesamt} und σG_{dual}). Ist die Summe aus durchschnittlicher Proliferation einer Zelllinie nach Behandlung mit dualen EGFR-BRD4-Inhibitoren ($\emptyset G_{dual}$) und entsprechender Standardabweichung (σG_{dual}) kleiner als die Differenz aus der durchschnittlichen Proliferation dieser Zelllinie nach Behandlung mit nicht auf BRD4 aktiven EGFR-Inhibitoren ($\emptyset G_{EGFR}$) und entsprechender Standardabweichung (σG_{EGFR}), so wurde die Zelllinie als sensitiv gegenüber einer duale EGFR-BRD4-Inhibition ermittelt. Besonders sensitive Zelllinien gegenüber dualer EGFR-BRD4-Inhibition sind Zelllinien, deren Summe aus durchschnittlicher Proliferation der Zelllinie nach Behandlung mit dualen EGFR-BRD4-Inhibitioren ($\emptyset G_{dual}$) und entsprechender Standardabweichung (σG_{dual}) kleiner ist als die Differenz aus der durchschnittlichen Proliferation dieser Zelllinie nach Behandlung mit EGFR-Inhibitoren ($\emptyset G_{gesamt}$) und entsprechender Standardabweichung (σG_{gesamt}). Die Berechnung ist in Gleichung IV.2.13 gegeben.

$$
\begin{aligned}
G_{dual,max} &= \emptyset G_{dual} + \sigma G_{dual} \\
G_{EGFR,min} &= \emptyset G_{EGFR} - \sigma G_{EGFR} \\
G_{gesamt,min} &= \emptyset G_{gesamt} - \sigma G_{gesamt} \\
Sensitivität(G_{dual,max}) &= \begin{cases} \text{insensitiv} & \text{für } G_{dual,max} \geq G_{EGFR,min} \\ \text{sensitiv} & \text{für } G_{gesamt,min} \leq G_{dual,max} < G_{EGFR,min} \\ \text{besonders sensitiv} & \text{für } G_{dual,max} < G_{gesamt,min} \end{cases}
\end{aligned}
\tag{IV.2.13}
$$

IV.2.17. Analyse und Vergleich von bezüglich EGFR-BRD4-Inhibition sensitiven und insensitven Zelllinien

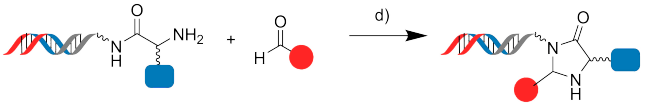
Mithilfe der CCLE (*Cancer Cell Line Encyclopedia* 2018; Cancer Cell Line Encyclopedia Consortium und Genomics of Drug Sensitivity in Cancer Consortium 2015) wurden die Mutationen in EGFR-BRD4-Inhibition sensitiven und insensitiven Zelllinien ermittelt und anschließend verglichen. Dazu wurden nur die besonders sensitiven Zelllinien (H647, HCC1195, H1975, H1563, H3122 und H2228) als sensitive Zelllinien verwendet. Die Zelllinien: HCC461, H2347, H1993, HeLa, H157, H322M und hPBMC wurden nicht verwendet, da sie entweder nicht besonders sensitiv sind oder nicht in der CCLE gefunden werden konnten. Damit ergaben sich folgende insensitive Zelllinien für die Analyse: H358, H2009, H441, H2030, HOP62, H2291, SW900, LCLC97TM1, DV90, H1944, HCC515, Calu6, H2122, HCC44, SKLU1, A427, A549, H2887, H1792, H460, H1155, H23, HCC15, H1915, H1299, H2087, H1395, H1666, H1755, HCC364, H1437, HCC366, H2286, H1581, H2077, H1781, H1819, H3255, HCC827, H596, H1703, Kelly, H2110, H520, HCC78, HCC1833, H2126, H292, H1838, HCC1359, HCC2429, H2172, H28, H2052, H838, HCC95, LCLC103H, Colo 699, H2085, H1568, EKVX, H2882, H661, SKMES1, H1693, HCC2450, H1793, H1648 und H322.

IV.2.18. Erstellung der Molekülfragmente für den Shape-Vergleich

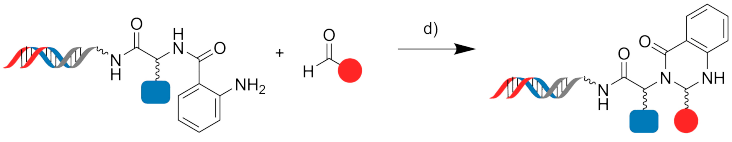
Alle Verbindungen in der CSD, welche mindestens ein Kohlenstoffatom besitzen, wurden von Sebastian Weigang heruntergeladen. Anschließend wurde zur Erstellung der Molekülfragmente ein Python-Skript von Dr. Hitesh Patel, welches RDKit (*RDKit: Open-source cheminformatics* 2018) verwendet, angewendet. Es wurden einerseits Fragmente, welche nur das zentrale Grundgerüst enthalten und andererseits Fragmente, die neben dem zentralen Grundgerüst je Substituent ein Atom als *Substituenten-Dummy* umfassen, erzeugt. Anschließend wurden doppelte Fragmente entfernt.

IV.2.19. *In silico* Erstellung der DNA-kodierten Bibliotheken

Die Edukte entstammen der ZINC-Datenbank (Irwin und Shoichet 2005; Sterling und Irwin 2015), welche nach Molekülmasse ($\leq$ 150 Da) und PAINS (Baell und Holloway 2010) mittels KNIME (Berthold *et al.* 2008) gefiltert wurde. Zudem wurden die Edukte mit allen natürlichen Aminosäuren ergänzt. Aus diesen Edukten wurden dann 50 (*kleine Bibliotheken*) bzw. 500 (*große Bibliotheken*) diverse, mit den für die jeweilige Reaktion benötigten Funktionalitäten ausgestattete, Edukte mittels des KNIME-Knotens *RDKit Diversity Picker* selektiert. Anschließend wurde die Reaktion *in silico* mithilfe des KNIME-Knotens *RDKit Two Component Reaction* durchgeführt. In der Abbildung IV.2.3 sind zwei verwendete DEL-kompatible Reaktionen dargestellt.



(a) DEL 7



(b) DEL 8

Abbildung IV.2.3.: Ausgewählte DEL-kompatible Reaktionen: Reaktion 7 (a) und Reaktion 8 (b). d: Heterozyklenbildung.

IV.2.20. Shape-Vergleich

Für den Shape-Vergleich wurden zwei verschiedene Methoden verwendet, einerseits FastROCS von OpenEye (*FastROCS Toolkit* 2018) und andererseits NPR nach Sauer und Schwarz (2003). FastROCS betrachtet das Volumen und die physikochemischen Eigenschaften der Moleküle (sogenannte Color-Features) als Color-Atoms mittels einer Gauß-Funktion. Das Volumen steht für den Shape. Zur Berechnung der Color-Atoms wird das implizite Mills Dean Color Kraftfeld verwendet und es werden z. B. Akzeptoren, Donoren, hydrophobe Bereiche, Ringe, An- und Kationen berücksichtigt (*Theory Toolkits* 2018). Der initiale Versuch erfolgte mithilfe von FastROCS (Python-Skripte: ShapeDatabaseClient.py und ShapeDatabaseServer.py, *FastROCS Toolkit* (2018)) und einem Tanimoto-Combo Schwellenwert von 1,7. Der Tanimoto-Combo setzt sich aus einem Shape-Tanimoto und einem Color-Tanimoto zusammen. Letzterer bewertet die Überlappung gleicher physikochemischer Eigenschaften. Zur Erstellung der Matrix wurde das Python-Skript *ShapeSimilarityMatrix.py* von OpenEye (*FastROCS Toolkit* 2018) benutzt. In der resultierenden Matrix wird die Distanz basierend auf dem Tanimoto-Combo angegeben.

Die NPR-Werte werden durch die Gleichungen IV.2.14 und IV.2.15 ermittelt. Dabei sind I_1, I_2 und I_3 das erste, zweite bzw. dritte Diagonalelement des diagonalisierten Trägheitstensors. Durch die Normalisierung ist die Beschreibung des Shapes eines Moleküls mittels NPR unabhängig von der Größe des Moleküls. Durch die Auftragung von npr1 auf die Abszisse und npr2 auf die Ordinate mittels Gnuplot 5.0 (*Gnuplot 5.0: An interactive plotting program* 2018) wird ein dreieckiger Raum aufgespannt. In der linken oberen Ecke befinden sich vornehmlich lineare stabförmige Moleküle, in der rechten oberen Ecke sind Moleküle mit einem sphärischen Shape zu finden und in der mittleren unteren Ecke sind flache scheibenförmige Moleküle angesiedelt (Sauer und Schwarz 2003). Die *principal moments of inertia* (Hauptträgheitsmomente) (PMI) und NPR-Deskriptoren wurden mithilfe von MOE 2016 berechnet. Dazu erfolgte zuvor die Erstellung der dreidimensionalen Struktur unter Verwendung des Kraftfeldes MMFF94x in MOE 2016 mit der Funktion *Rebuild 3D* und den Optionen *Dominant Protonation*, *Preserve Existing Chirality* und *Scale to Reasonable Bond Length*.

$$npr1 = \frac{I_1}{I_3} \tag{IV.2.14}$$

$$npr2 = \frac{I_2}{I_3} \tag{IV.2.15}$$

Ein Unterschied zwischen FastROCS und NPR ist, dass FastROCS die Shape-Ähnlichkeit immer in Relation zu einem gegebenen Referenz-Shape kalkuliert, während NPR unabhängig von einer Referenz ist.

IV.2.21. Erstellung der Bioisostere

Aus der ChEMBL Version 20 (Bento *et al.* 2014) wurden mittels Pipeline Pilot 9.1 (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015) und SQL-Befehlen alle Einträge der ChEMBL herausgefiltert, (i) die nicht „intermediate“ kuriert sind, (ii) die eine Aktivitätsstandardbeziehung von „=“ haben, (iii) deren Molekülmasse nicht größer als 1000 Da ist, (iv) deren Konfidenzscore größer als sieben ist, (v) deren Aktivitätsstandardtyp entweder „IC_{50}“, „EC_{50}“, „K_i“ oder „K_d“ ist und (vi) die einen Aktivitätsstandardwert von $<$ 10 μM besitzen. Auf der Basis dieser Selektion der ChEMBL wurden dann in einem Pipeline Pilot Workflow die MMPs mithilfe der Komponente *Matched Molecular Pairs* generiert. Dazu wurde unter *Compound ID Property* und *Activity Property* das Stichwort, unter dem die Verbindungs-ID respektive die Aktivität vermerkt ist, angegeben. Alle MMPs, deren Bioaktivität sich um weniger als eine Größenordnung unterschied, wurden als initiale Bioisostere deklariert. Anschließend wurde je bioisosterem Paar nur ein Beispiel beibehalten (einzigartige bioisostere Paare) und die Verlässlichkeit der beobachteten Ähnlichkeit in der Bioaktivität bewertet. Beide Schritte erfolgten durch entsprechende Python-Skripte. Zur Bewertung der Verlässlichkeit wurde die Anzahl an bioisosteren Austauschen je bioisosterem Paar ermittelt, in denen die Bioaktivität beibehalten wurde und in denen eine Änderung der Bioaktivität um mehr als eine Größenordnung beobachtet wurde. Erfolgten mindestens 75 % der Austausche für ein bioisosteres Paar unter Beibehaltung der Bioaktivität, so wurde dieses Paare als zuverlässiges bioisosteres Paar in Betracht gezogen. Zudem wurden nur bioisostere Paare in die *zuverlässigen bioisosteren Paare* übernommen, welche mindestens 20 mal gefunden werden konnten.

Teil V.

Literaturverzeichnis

Literatur

Allen, B. K., S. Mehta, S. W. J. Ember, E. Schonbrunn, N. Ayad und S. C. Schürer (2015). „Large-scale computational screening identifies first in class multitarget inhibitor of EGFR kinase and BRD4“. In: *Scientific Reports* 5. DOI: 10.1038/srep16924.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers und D. J. Lipman (1990). „Basic local alignment search tool“. In: *Journal of Molecular Biology* 215.3, S. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Aouali, N., A. Broukou, M. Bosseler, O. Keunen, V. Schlesser, B. Janji, V. Palissot, P. Stordeur und G. Berchem (2015). „Epigenetic activity of peroxisome proliferator-activated receptor gamma agonists inreases the anticancer effect of histone deacetylase inhibitors on multiple myeloma cells“. In: *PLoS ONE* 10.6, S. 1–17. DOI: 10.1371/journal.pone.0130339.

Arrowsmith, C. H., J. E. Audia, C. Austin, J. Baell, J. Bennett, J. Blagg, C. Bountra, P. E. Brennan, P. J. Brown, M. E. Bunnage, C. Buser-Doepner, R. M. Campbell, A. J. Carter, P. Cohen, R. A. Copeland, B. Cravatt, J. L. Dahlin, D. Dhanak, A. M. Edwards, M. Frederiksen, S. V. Frye, N. Gray, C. E. Grimshaw, D. Hepworth, T. Howe, K. V. M. Huber, J. Jin, S. Knapp, J. D. Kotz, R. G. Kruger, D. Lowe, M. M. Mader, B. Marsden, A. Mueller-Fahrnow, S. Müller, R. C. O'Hagan, D. R. Overington John P Owen, S. H. Rosenberg, R. Ross, B. Roth, M. Schapira, S. L. Schreiber, B. Shoichet, M. Sundström, G. Superti-Furga, J. Taunton, L. Toledo-Sherman, C. Walpole, M. A. Walters, T. M. Willson, P. Workman, R. N. Young und W. J. Zuercher (2015). „The promise and peril of chemical probes“. In: *Nature Chemical Biology* 11, S. 536–541. DOI: 10.1038/nchembio.1867.

Backman, T. W. H., D. S. Evans und T. Girke (2017). „Large-scale bioactivity analysis of the small-molecule assayed proteome“. In: *PLoS ONE* 12.2, S. 1–36. DOI: 10.1371/journal.pone.0171413.

Baell, J. B. (2016). „Feeling nature's PAINS: Natural products, natural product drugs, and pan assay interference compounds (PAINS)“. In: *Journal of Natural Products* 79.3, S. 616–628. DOI: 10.1021/acs.jnatprod.5b00947.

Baell, J. B. und G. A. Holloway (2010). „New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays“. In: *Journal of Medicinal Chemistry* 53.7, S. 2719–2740. DOI: 10.1021/jm901137j.

Bajusz, D., A. Rácz und K. Héberger (2015). „Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?“ In: *Journal of Cheminformatics* 7.1, S. 20. DOI: 10.1186/s13321-015-0069-3.

Balamurugan, R., F. J. Dekker und H. Waldmann (2005). „Design of compound libraries based on natural product scaffolds and protein structure similarity clustering (PSSC)“. In: *Molecular BioSystems* 1.1, S. 36–45. DOI: 10.1039/B503623B.

Ballatore, C., D. M. Huryn und A. B. Smith (2013). „Carboxylic acid (bio)isosteres in drug design“. In: *ChemMedChem* 8.3, S. 385–395. DOI: 10.1002/cmdc.201200585.

Baxter, C. A., C. W. Murray, D. E. Clark, D. R. Westhead und M. D. Eldridge (1998). „Flexible docking using tabu search and an empirical estimate of binding affinity“. In: *Proteins: Structure, Function, and Bioinformatics* 33.3, S. 367–382. DOI: 10.1002/(SICI)1097-0134(19981115)33:3<367::AID-PROT6>3.0.CO;2-W.

Beitz, E. (2000). „TeXshade: Shading and labeling of multiple sequence alignments using LaTeX2e“. In: *Bioinformatics* 16.2, S. 135–139. DOI: 10.1093/bioinformatics/16.2.135.

Bento, A. P., A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos und J. P. Overington (2014). „The ChEMBL bioactivity database: An update“. In: *Nucleic Acids Research* 42.D1, S. D1083–D1090. DOI: 10.1093/nar/gkt1031.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov und P. E. Bourne (2000). „The protein data bank“. In: *Nucleic Acids Research* 28.1, S. 235–242. DOI: 10.1093/nar/28.1.235.

Berthold, M. R., N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel und B. Wiswedel (2008). „KNIME: The konstanz information miner“. In: *Data Analysis, Machine Learning and Applications*. Hrsg. von C. reisach, H. Burkhardt, L. Schmidt-Thieme und R. Decker. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 319–326. ISBN: 978-3-540-78246-9.

Bhagwat, A. S., J.-S. Roe, B. Y. L. Mok, A. F. Hohmann, J. Shi und C. R. Vakoc (2016). „BET bromodomain inhibition releases the mediator complex from select cis-regulatory elements“. In: *Cell Reports* 15.3, S. 519–530. DOI: 10.1016/j.celrep.2016.03.054.

BIOVIA Pipeline Pilot: BIOVIA Workbook (2015). San Diego: Dassault Systèmes.

Bolin, S., A. Borgenvik, C. U. Persson, A. Sundström, J. Qi, J. E. Bradner, W. A. Weiss, Y.-J. Cho, H. Weishaupt und F. J. Swartling (2018). „Combined BET bromodomain and CDK2 inhibition in MYC-driven medulloblastoma“. In: *Oncogene* 37.21, S. 2850–2862. DOI: 10.1038/s41388-018-0135-1.

Bonandi, E., M. S. Christodoulou, G. Fumagalli, D. Perdicchia, G. Rastelli und D. Passarella (2017). „The 1,2,3-triazole ring as a bioisostere in medicinal chemistry“. In: *Drug Discovery Today* 22.10, S. 1572–1581. DOI: doi.org/10.1016/j.drudis.2017.05.014.

Brägelmann, J., M. A. Dammert, F. Dietlein, J. M. Heuckmann, A. Choidas, S. Böhm, A. Richters, D. Basu, V. Tischler, C. Lorenz, P. Habenberger, Z. Fang, S. Ortiz-Cuaran, F. Leenders, J. Eickhoff, U. Koch, M. Getlik, M. Termathe, M. Sallouh, Z. Greff, Z. Varga, H. Balke-Want, C. A. French, M. Peifer, H. C. Reinhardt, L. Örfi, G. Kéri, S. Ansén, L. C. Heukamp, R. Büttner, D. Rauh, B. M. Klebl, R. K. Thomas und M. L. Sos (2017). „Systematic kinase inhibitor profiling identifies CDK9 as a synthetic lethal target in NUT midline carcinoma“. In: *Cell Reports* 20.12, S. 2833–2845. DOI: 10.1016/j.celrep.2017.08.082.

Breinbauer, R., I. R. Vetter und H. Waldmann (2002). „From protein domains to drug candidates—natural products as guiding principles in the design and synthesis of compound libraries“. In: *Angewandte Chemie International Edition* 41.16, S. 2878–2890. DOI: 10.1002/1521-3773(20020816)41:16<2878::AID-ANIE2878>3.0.CO;2-B.

Brown, A. C. und T. R. Fraser (1868). „On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia“. In: *Journal of Anatomy and Physiology* 2.2, S. 224–242.

Brown, J. D., Z. B. Feldman, S. P. Doherty, J. M. Reyes, P. B. Rahl, C. Y. Lin, Q. Sheng, Q. Duan, A. J. Federation, A. L. Kung, S. M. Haldar, R. A. Young, J. Plutzky und J. E. Bradner (2018). „BET bromodomain proteins regulate enhancer function during adipogenesis“. In: *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.1711155115.

Brown, J. D., C. Y. Lin, Q. Duan, G. Griffin, A. Federation, R. M. Paranal, S. Bair, G. Newton, A. Lichtman, A. Kung, T. Yang, H. Wang, F. W. Luscinskas, K. Croce, J. E. Bradner und J. Plutzky (2014). „NF-κB directs dynamic super enhancer formation in inflammation and atherogenesis“. In: *Molecular Cell* 56.2, S. 219–231. DOI: 10.1016/j.molcel.2014.08.024.

Bunke, H. und K. Shearer (1998). „A graph distance metric based on the maximal common subgraph“. In: *Pattern Recognition Letters* 19.3, S. 255–259. DOI: 10.1016/S0167-8655(97)00179-7.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer und T. L. Madden (2009). „BLAST+: Architecture and applications“. In: *BMC Bioinformatics* 10, S. 421. DOI: 10.1186/1471-2105-10-421.

Cancer Cell Line Encyclopedia Consortium und Genomics of Drug Sensitivity in Cancer Consortium (2015). „Pharmacogenomic agreement between two cancer cell line data sets“. In: *Nature* 528.7580, S. 84–87. DOI: 10.1038/nature15736.

Carlino, L. und G. Rastelli (2016). „Dual kinase-bromodomain inhibitors in anticancer drug discovery: A structural and pharmacological perspective“. In: *Journal of Medicinal Chemistry* 59.20, S. 9305–9320. DOI: 10.1021/acs.jmedchem.6b00438.

Carta, G., A. J. S. Knox und D. G. Lloyd (2007). „Unbiasing scoring functions: A new normalization and rescoring strategy“. In: *Journal of Chemical Information and Modeling* 47.4, S. 1564–1571. DOI: 10.1021/ci600471m.

Chartier, M., E. Adriansen und R. Najmanovich (2016). „IsoMIF Finder: Online detection of binding site molecular interaction field similarities“. In: *Bioinformatics* 32.4, S. 621–623. DOI: 10.1093/bioinformatics/btv616.

Chartier, M. und R. Najmanovich (2015). „Detection of binding site molecular interaction field similarities“. In: *Journal of Chemical Information and Modeling* 55.8, S. 1600–1615. DOI: 10.1021/acs.jcim.5b00333.

Chen, V. B., W. B. Arendall III, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson und D. C. Richardson (2010). „*MolProbity*: All-atom structure validation for macromolecular crystallography“. In: *Acta Crystallographica Section D* 66.1, S. 12–21. DOI: 10.1107/S0907444909042073.

Chevillard, F. und P. Kolb (2015). „SCUBIDOO: A large yet screenable and easily searchable database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability“. In: *Journal of Chemical Information and Modeling* 55.9, S. 1824–1835. DOI: 10.1021/acs.jcim.5b00203.

Dang, C. V., E. P. Reddy, K. M. Shokat und L. Soucek (2017). „Drugging the „undruggable“ cancer targets“. In: *Nature Reviews Cancer* 17, S. 502–508. DOI: 10.1038/nrc.2017.36.

DeGoey, D. A., H.-J. Chen, P. B. Cox und M. D. Wendt (2017). „Beyond the rule of 5: Lessons learned from AbbVie's drugs and compound collection“. In: *Journal of Medicinal Chemistry*. DOI: 10.1021/acs.jmedchem.7b00717.

Delmore, J. E., G. C. Issa, M. E. Lemieux, P. B. Rahl, J. Shi, H. M. Jacobs, E. Kastritis, T. Gilpatrick, R. M. Paranal, J. Qi, M. Chesi, A. C. Schinzel, M. R. McKeown, T. P. Heffernan, C. R. Vakoc, P. L. Bergsagel, I. M. Ghobrial, P. G. Richardson, R. A. Young, W. C. Hahn, K. C. Anderson, A. L. Kung, J. E. Bradner und C. S. Mitsiades (2011). „BET bromodomain inhibition as a therapeutic strategy to target c-Myc“. In: *Cell* 146.6, S. 904–917. DOI: 10.1016/j.cell.2011.08.017.

Denis, G. V., B. S. Nikolajczyk und G. R. Schnitzler (2010). „An emerging role for bromodomain-containing proteins in chromatin regulation and transcriptional control of adipogenesis“. In: *FEBS Letters* 584.15, S. 3260–3268. DOI: 10.1016/j.febslet.2010.05.030..

Desaphy, J., B. Guillaume, R. Didier und E. Kellenberger (2015). „Sc-PDB: A 3D-database of ligandable binding sites - 10 years on“. In: *Nucleic Acids Research* 43.D1, S. D399–D404.

Dimova, D., K. Heikamp, D. Stumpfe und J. Bajorath (2013). „Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets“. In: *Journal of Medicinal Chemistry* 56.8, S. 3339–3345. DOI: 10.1021/jm400147j.

Dittmann, A., T. Werner, C.-W. Chung, M. M. Savitski, M. Fälth Savitski, P. Grandi, C. Hopf, M. Lindon, G. Neubauer, R. K. Prinjha, M. Bantscheff und G. Drewes (2014). „The commonly used PI3-Kinase probe LY294002 is an inhibitor of BET bromodomains“. In: *ACS Chemical Biology* 9.2, S. 495–502. DOI: 10.1021/cb400789e.

Doak, B. C., J. Zheng, D. Dobritzsch und J. Kihlberg (2016). „How beyond rule of 5 drugs and clinical candidates bind to their targets“. In: *Journal of Medicinal Chemistry* 59.6, S. 2312–2327. DOI: 10.1021/acs.jmedchem.5b01286.

Dömling, A., W. Wang und K. Wang (2012). „Chemistry and biology Of multicomponent reactions“. In: *Chemical Reviews* 112.6, S. 3083–3135. DOI: 10.1021/cr100233r.

Droschinsky, A., N. M. Kriege und P. Mutzel (2016). „Faster algorithms for the maximum common subtree isomorphism problem“. In: *41st International Symposium on Mathematical Foundations of Computer Science (MFCS 2016)*. Hrsg. von P. Faliszewski, A. Muscholl und R. Niedermeier. Bd. 58. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 33:1–33:14. ISBN: 978-3-95977-016-3. DOI: 10.4230/LIPIcs.MFCS.2016.33.

— (2018). „Largest weight common subtree embeddings with distance penalties“. In: *43rd International Symposium on Mathematical Foundations of Computer Science (MFCS 2018)*. Hrsg. von I. Potapov, P. Spirakis und J. Worrell. Bd. 117. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 54:1–54:15. ISBN: 978-3-95977-086-6.

Droschinsky, A., N. Kriege und P. Mutzel (2017). „Finding largest common substructures of molecules in quadratic time“. In: *43rd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2017): Theory and Practice of Computer Science*. Hrsg. von B. Steffen, C. Baier, M. van den Brand, J. Eder, M. Hinchey und T. Margaria. Cham: Springer International Publishing, S. 309–321. ISBN: 978-3-319-51963-0.

Duvenaud, D., D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik und R. P. Adams (2015). „Convolutional networks on graphs for learning molecular fingerprints“. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Montreal, Canada: MIT Press, S. 2224–2232.

Echevarría-Vargas, I. M., P. I. Reyes-Uribe, A. N. Guterres, X. Yin, A. V. Kossenkov, Q. Liu, G. Zhang, C. Krepler, C. Cheng, Z. Wei, R. Somasundaram, G. Karakousis, W. Xu, J. J. Morrissette, Y. Lu, G. B. Mills, R. J. Sullivan, M. Benchun, D. T. Frederick, G. Boland, K. T. Flaherty, A. T. Weeraratna, M. Herlyn, R. Amaravadi, L. M. Schuchter, C. E. Burd, A. E. Aplin, X. Xu und J. Villanueva (2018). „Co-targeting BET and MEK as salvage therapy for MAPK and checkpoint inhibitor-resistant melanoma“. In: *EMBO Molecular Medicine* 10.5. DOI: 10.15252/emmm.201708446.

Ehrt, C., T. Brinkjost und O. Koch (2018). „A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs)“.

In: *PLOS Computational Biology* 14.11, S. 1–50. DOI: 10.1371/journal.pcbi.1006483.

El Akoum, S. (2014). „PPAR gamma at the crossroads of health and disease: A masterchef in metabolic homeostasis“. In: *Endocrinology & Metabolic Syndrome* 03.01. DOI: 10.4172/2161-1017.1000126.

Eldridge, M. D., C. W. Murray, T. R. Auton, G. V. Paolini und R. P. Mee (1997). „Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes“. In: *Journal of Computer-Aided Molecular Design* 11.5, S. 425–445. DOI: 10.1023/A:1007996124545.

Ellis, D., J. Furner-Hines und P. Willett (1993). „Measuring the degree of similarity between objects in text retrieval systems“. In: *Perspectives in Information Management* 3.2, S. 128–149.

Ember, S. W. J., J.-Y. Zhu, S. H. Olesen, M. P. Martin, A. Becker, N. Berndt, G. I. Georg und E. Schönbrunn (2014). „Acetyl-lysine binding site of bromodomain-containing protein 4 (BRD4) interacts with diverse kinase inhibitors“. In: *ACS Chemical Biology* 9.5, S. 1160–1171. DOI: 10.1021/cb500072z.

Ember, S. W., Q. T. Lambert, N. Berndt, S. Gunawan, M. Ayaz, M. Tauro, J.-Y. Zhu, P. J. Cranfill, P. Greninger, C. C. Lynch, C. H. Benes, H. R. Lawrence, G. W. Reuther, N. J. Lawrence und E. Schönbrunn (2017). „Potent dual BET bromodomain-kinase inhibitors as value-added multitargeted chemical probes and cancer therapeutics“. In: *Molecular Cancer Therapeutics* 16.6, S. 1054–1067. DOI: 10.1158/1535-7163.MCT-16-0568-T.

Emsley, P. und K. Cowtan (2004). „Coot: Model-building tools for molecular graphics“. In: *Acta Crystallographica Section D* 60.12 Part 1, S. 2126–2132. DOI: 10.1107/S0907444904019158.

Engel, J. (2018). „Strukturbasiertes Design und Entwicklung kovalenter Kinaseinhibitoren zur Umgehung der durch EGFR-T790M vermittelten Wirkstoffresistenz im nichtkleinzelligen Lungenkarzinom“. Dissertation. Fakultät für Chemie und Chemische Biologie, TU Dortmund.

Engel, J., C. Becker, J. Lategahn, M. Keul, J. Ketzer, T. Mühlenberg, L. Kollipara, C. Schultz-Fademrecht, R. P. Zahedi, S. Bauer und D. Rauh (2016). „Insight into the inhibition of drug-resistant mutants of the receptor yrosine kinase EGFR“. In: *Angewandte Chemie International Edition* 55.36, S. 10909–10912. DOI: 10.1002/anie.201605011.

Engel, J., J. Lategahn und D. Rauh (2016). „Hope and disappointment: Covalent inhibitors to overcome drug resistance in non-small cell lung cancer“. In: *ACS Medicinal Chemistry Letters* 7.1, S. 2–5. DOI: 10.1021/acsmedchemlett.5b00475.

Engel, J., A. Richters, M. Getlik, S. Tomassi, M. Keul, M. Termathe, J. Lategahn, C. Becker, S. Mayer-Wrangowski, C. Grütter, N. Uhlenbrock, J. Krüll, N. Schaumann, S. Eppmann, P. Kibies, F. Hoffgaard, J. Heil, S. Menninger, S. Ortiz-Cuaran, J. M. Heuckmann, V. Tinnefeld, R. P. Zahedi, M. L. Sos, C. Schultz-Fademrecht, R. K.

Thomas, S. M. Kast und D. Rauh (2015). „Targeting drug resistance in EGFR with covalent inhibitors: A structure-based design approach“. In: *Journal of Medicinal Chemistry* 58.17, S. 6844–6863. DOI: 10.1021/acs.jmedchem.5b01082.

Engel, J., S. Smith, J. Lategahn, H. L. Tumbrink, L. Goebel, C. Becker, E. Hennes, M. Keul, A. Unger, H. Müller, M. Baumann, C. Schultz-Fademrecht, G. Günther, J. G. Hengstler und D. Rauh (2017). „Structure-guided development of covalent and mutant-selective pyrazolopyrimidines to target T790M drug resistance in epidermal growth factor receptor“. In: *Journal of Medicinal Chemistry* 60.18, S. 7725–7744. DOI: 10.1021/acs.jmedchem.7b00515.

Engelke, C. G. und A. M. Chinnaiyan (2015). „aBETting therapeutic resistance by Wnt signaling“. In: *Cell Research* 25, S. 1178–1188. DOI: 10.1038/cr.2015.127.

Ertl, P. und A. Schuffenhauer (2009). „Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions“. In: *Journal of Cheminformatics* 1.1, S. 8. DOI: 10.1186/1758-2946-1-8.

Evans, B. E., K. E. Rittle, M. G. Bock, R. M. DiPardo, R. M. Freidinger, W. L. Whitter, G. F. Lundell, D. F. Veber, P. S. Anderson, R. S. L. Chang, V. J. Lotti, D. J. Cerino, T. B. Chen, P. J. Kling, K. A. Kunkel, J. P. Springer und J. Hirshfield (1988). „Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists“. In: *Journal of Medicinal Chemistry* 31.12, S. 2235–2246. DOI: 10.1021/jm00120a002.

Ferri, E., C. Petosa und C. E. McKenna (2016). „Bromodomains: Structure, function and pharmacology of inhibition“. In: *Biochemical Pharmacology* 106, S. 1–18. DOI: 10.1016/j.bcp.2015.12.005.

Filippakopoulos, P. und S. Knapp (2014). „Targeting bromodomains: Epigenetic readers of lysine acetylation“. In: *Nature Reviews Drug Discovery* 13.5, S. 337–356. DOI: 10.1038/nrd4286.

Filippakopoulos, P., S. Picaud, M. Mangos, T. Keates, J.-P. Lambert, D. Barsyte-Lovejoy, I. Felletar, R. Volkmer, S. Müller, T. Pawson, A.-C. Gingras, C. H. Arrowsmith und S. Knapp (2012). „Histone recognition and large-scale structual analysis of the human bromodomain family“. In: *Cell* 149.1, S. 214–231. DOI: 10.1016/j.cell.2012.02.013.

Filippakopoulos, P., J. Qi, S. Picaud, Y. Shen, W. B. Smith, O. Fedorov, E. M. Morse, T. Keates, T. T. Hickman, I. Felletar, M. Philpott, S. Munro, M. R. McKeown, Y. Wang, A. L. Christie, N. West, M. J. Cameron, B. Schwartz, T. D. Heightman, N. La Thangue, C. A. French, O. Wiest, A. L. Kung, S. Knapp und J. E. Bradner (2010). „Selective inhibition of BET bromodomains“. In: *Nature* 468.7327, S. 1067–1073. DOI: 10.1038/nature09504.

Friedrich, T., B. Richter, T. Gaiser, C. Weiss, K.-P. Janssen, H. Einwächter, R. M. Schmid, M. P. Ebert und E. Burgermeister (2013). „Deficiency of caveolin-1 in Apcmin/+ mice promotes colorectal tumorigenesis“. In: *Carcinogenesis* 34.9, S. 2109–2118. DOI: 10.1093/carcin/bgt142.

Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel und A. Bairoch (2005). „Protein identification and analysis tools on the ExPASy server“. In: *The Proteomics Protocols Handbook*. Hrsg. von J. M. Walker. Humana Press. Kap. 50, S. 571–607.

Gaulton, A., A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit und A. R. Leach (2017). „The ChEMBL database in 2017“. In: *Nucleic Acids Research* 45.D1, S. D945–D954. DOI: 10.1093/nar/gkw1074.

Gehde, R. (2018). „Erweiterung von Pharmakophor-basierten Fingerprints mittels Randomisierung“. Bachelorarbeit. Fakultät für Informatik, TU Dortmund.

Getlik, M. (2011). „Strukturbasierte Entwicklung und Synthese von Enzyminhibitoren“. Dissertation. Fakultät Chemie, TU Dortmund.

Godden, J. W., L. Xue und J. Bajorath (2000). „Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients“. In: *Journal of Chemical Information and Computer Sciences* 40.1, S. 163–166. DOI: 10.1021/ci990316u.

Günther, M., J. Lategahn, M. Juchum, E. Döring, M. Keul, J. Engel, H. L. Tumbrink, D. Rauh und S. Laufer (2017). „Trisubstituted pyridinylimidazoles as potent inhibitors of the clinically resistant L858R/T790M/C797S EGFR mutant: Targeting of both hydrophobic regions and the phosphate binding site“. In: *Journal of Medicinal Chemistry* 60.13, S. 5613–5637. DOI: 10.1021/acs.jmedchem.7b00316.

Harrison, R. K. (2016). „Phase II and phase III failures: 2013–2015“. In: *Nature Reviews Drug Discovery* 15.12, S. 817–818. DOI: 10.1038/nrd.2016.184.

Hartenfeller, M., M. Eberle, P. Meier, C. Nieto-Oberhuber, K.-H. Altmann, G. Schneider, E. Jacoby und S. Renner (2011). „A collection of robust organic synthesis reactions for in silico molecule design“. In: *Journal of Chemical Information and Modeling* 51.12, S. 3093–3098. DOI: 10.1021/ci200379p.

Herndon, W. C. und S. H. Bertz (1987). „Linear notations and molecular graph similarity“. In: *Journal of Computational Chemistry* 8.4, S. 367–374. DOI: 10.1002/jcc.540080413.

Horváth, T., J. Ramon und S. Wrobel (2010). „Frequent subgraph mining in outerplanar graphs“. In: *Data Mining and Knowledge Discovery* 21.3, S. 472–508. DOI: 10.1007/s10618-009-0162-1.

Hu, Y.-J., H. Belaghzal, W.-Y. Hsiao, J. Qi, J. E. Bradner, D. A. Guertin, S. Sif und A. N. Imbalzano (2015). „Transcriptional and post-transcriptional control of adipocyte differentiation by Jumonji domain-containing protein 6“. In: *Nucleic Acids Research* 43.16, S. 7790–7804. DOI: 10.1093/nar/gkv645.

Hu, Y., D. Stumpfe und J. Bajorath (2017). „Recent advances in scaffold hopping“. In: *Journal of Medicinal Chemistry* 60.4, S. 1238–1246. DOI: 10.1021/acs.jmedchem.6b01437.

Humbeck, L. und O. Koch (2017). „What can we learn from bioactivity data? Chemoinformatics tools and applications in chemical biology research“. In: *ACS Chemical Biology* 12.1, S. 23–35. DOI: 10.1021/acschembio.6b00706.

Humbeck, L., S. Weigang, T. Schäfer, P. Mutzel und O. Koch (2018). „CHIPMUNK: A virtual synthesizable small-molecule library for medicinal chemistry, exploitable for protein–protein interaction modulators“. In: *ChemMedChem* 13.6, S. 532–539. DOI: 10.1002/cmdc.201700689.

Hussong, M., S. T. Börno, M. Kerick, A. Wunderlich, A. Franz, H. Sültmann, B. Timmermann, H. Lehrach, M. Hirsch-Kauffmann und M. R. Schweiger (2014). „The bromodomain protein BRD4 regulates the KEAP1/NRF2-dependent oxidative stress response“. In: *Cell Death & Disease* 5, e1195. DOI: 10.1038/cddis.2014.157.

Irwin, J. J., D. Duan, H. Torosyan, A. K. Doak, K. T. Ziebart, T. Sterling, G. Tumanian und B. K. Shoichet (2015). „An aggregation advisor for ligand discovery“. In: *Journal of Medicinal Chemistry* 58.17, S. 7076–7087. DOI: 10.1021/acs.jmedchem.5b01105.

Irwin, J. J. und B. K. Shoichet (2005). „ZINC–a free database of commercially available compounds for virtual screening“. In: *Journal of Chemical Information and Modeling* 45.1, S. 177–182. DOI: 10.1021/ci049714.

Ishay-Ronen, D., M. Diepenbruck, R. K. R. Kalathur, N. Sugiyama, S. Tiede, R. Ivanek, G. Bantug, M. F. Morini, J. Wang, C. Hess und G. Christofori (2019). „Gain fat-lose metastasis: Converting invasive breast cancer cells into adipocytes inhibits cancer metastasis“. In: *Cancer Cell* 35.1, S. 1535–6108. DOI: 10.1016/j.ccell.2018.12.002.

Itzen, F., A. K. Greifenberg, C. A. Bösken und M. Geyer (2014). „Brd4 activates P-TEFb for RNA polymerase II CTD phosphorylation“. In: *Nucleic Acids Research* 42.12, S. 7577–7590. DOI: 10.1093/nar/gku449.

Jadhav, A., R. S. Ferreira, C. Klumpp, B. T. Mott, C. P. Austin, J. Inglese, C. J. Thomas, D. J. Maloney, B. K. Shoichet und A. Simeonov (2010). „Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease“. In: *Journal of Medicinal Chemistry* 53.1, S. 37–51. DOI: 10.1021/jm901070c.

Jasper, J. B., L. Humbeck, T. Brinkjost und O. Koch (2018). „A novel interaction fingerprint derived from per atom score contributions: Exhaustive evaluation of interaction fingerprint performance in docking based virtual screening“. In: *Journal of Cheminformatics* 10.1, S. 15. DOI: 10.1186/s13321-018-0264-0.

Jerabek-Willemsen, M., T. André, R. Wanner, H. M. Roth, S. Duhr, P. Baaske und D. Breitsprecher (2014). „MicroScale Thermophoresis: Interaction analysis and beyond“. In: *Journal of Molecular Structure* 1077. Fluorescence studies of biomolecular association processes., S. 101–113. DOI: 10.1016/j.molstruc.2014.03.009.

Jones, G., P. Willett, R. C. Glen, A. R. Leach und R. Taylor (1997). „Development and validation of a genetic algorithm for flexible docking“. In: *Journal of Molecular Biology* 267.3, S. 727–748. DOI: 10.1006/jmbi.1996.0897.

Joosten, R. P., F. Long, G. N. Murshudov und A. Perrakis (2014). „The PDB_REDO server for macromolecular structure model optimization“. In: *IUCrJ* 1.Part 4, S. 213–220. DOI: 10.1107/S2052252514009324.

Kabsch, W. (2010). „XDS“. In: *Acta Crystallographica Section D* 66.2, S. 125–132. DOI: 10.1107/S0907444909047337.

Kellenberger, E., P. Muller, C. Schalon, G. Bret, N. Foata und D. Rognan (2006). „sc-PDB: An annotated database of druggable binding sites from the protein data bank“. In: *Journal of Chemical Information and Modeling* 46.2, S. 717–727. DOI: 10.1021/ci050372x.

Kersten, S., B. Desvergne und W. Wahli (2000). „Roles of PPARs in health and disease“. In: *Nature* 405.6785, S. 421–424. DOI: 10.1038/35013000.

Kitchen, D. B., H. Decornez, J. R. Furr und J. Bajorath (2004). „Docking and scoring in virtual screening for drug discovery: Methods and applications“. In: *Nature Reviews Drug Discovery* 3.11, S. 935–949. DOI: 10.1038/nrd1549.

Kleczko, E. K. und L. E. Heasley (2018). „Mechanisms of rapid cancer cell reprogramming initiated by targeted receptor tyrosine kinase inhibitors and inherent therapeutic vulnerabilities“. In: *Molecular Cancer* 17.1. DOI: 10.1186/s12943-018-0816-y.

Koch, M. A. und H. Waldmann (2005). „Protein structure similarity clustering and natural product structure as guiding principles in drug discovery“. In: *Drug Discovery Today* 10.7, S. 471–483. DOI: 10.1016/S1359-6446(05)03419-7.

Koch, M. A., L.-O. Wittenberg, S. Basu, D. A. Jeyaraj, E. Gourzoulidou, K. Reinecke, A. Odermatt und H. Waldmann (2004). „Compound library development guided by protein structure similarity clustering and natural product structure“. In: *Proceedings of the National Academy of Sciences* 101.48, S. 16721–16726. DOI: 10.1073/pnas.0404719101.

Koes, D. R., A. Dömling und C. J. Camacho (2018). „AnchorQuery: Rapid online virtual screening for small-molecule protein-protein interaction inhibitors“. In: *Protein Science* 27.1, S. 229–232. DOI: 10.1002/pro.3303.

Kotla, S. und G. N. Rao (2015). „Reactive oxygen species (ROS) mediate p300-dependent STAT1 protein interaction with peroxisome proliferator-activated receptor (PPAR)-γ in CD36 protein expression and foam cell formation“. In: *The Journal of Biological Chemistry* 290.51, S. 30306–30320. DOI: 10.1074/jbc.M115.686865.

Kriege, N. M., L. Humbeck und O. Koch (2019). „Chemical Similarity and Substructure Searches“. In: *Encyclopedia of Bioinformatics and Computational Biology*. Hrsg. von S. Ranganathan, M. Gribskov, K. Nakai und C. Schönbach. Oxford: Academic Press, S. 640–649. ISBN: 978-0-12-811432-2. DOI: 10.1016/B978-0-12-809633-8.20195-7.

Kumar, K. J., S. S. Raza, L. M. Knab, C. R. Chow, B. Kwok, D. J. Bentrem, R. Popovic, K. Ebine, J. D. Licht und H. G. Munshi (2015). „GLI2-dependent c-MYC upregulation mediates resistance of pancreatic cancer cells to the BET bromodomain inhibitor JQ1“. In: *Scientific Reports*. DOI: 10.1038/srep09489.

Kurimchak, A. M., C. Shelton, K. E. Duncan, K. J. Johnson, J. R. Brown, S. W. O'Brien, R. T. Gabbasov, L. S. Fink, Y. Li, N. T. Lounsbury, M. A. Abou-Gharbia, W. E. Childers, D. C. Connolly, J. Chernoff, J. R. Peterson und J. S. Duncan (2016). „Resistance to BET bromodomain inhibitors is mediated by kinome reprogramming in ovarian cancer". In: *Cell Reports* 16.5, S. 1273–1286. DOI: 10.1016/j.celrep.2016.06.091.

Labbé, C. M., M. A. Kuenemann, B. Zarzycka, G. Vriend, G. A. F. Nicolaes, D. Lagorce, M. A. Miteva, B. O. Villoutreix und O. Sperandio (2016). „iPPI-DB: An online database of modulators of protein-protein interactions". In: *Nucleic Acids Research* 44.D1, S. D542–7. DOI: 10.1093/nar/gkv982..

Larsen, B. und C. Aone (1999). „Fast and effective text mining using linear-time document clustering". In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '99. San Diego, California, USA: ACM, S. 16–22. ISBN: 1-58113-143-7. DOI: 10.1145/312129.312186.

Law, V., C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou und D. S. Wishart (2013). „DrugBank 4.0: Shedding new light on drug metabolism". In: *Nucleic Acids Research* 42.D1, S. D1091–D1097. DOI: 10.1093/nar/gkt1068.

Leach, A. R. und V. Gillet (2007). „Molecular Descriptors". In: *An introduction to chemoinformatics*. Dordrecht: Springer Netherlands, S. 53–74. ISBN: 978-1-4020-6291-9. DOI: 10.1007/978-1-4020-6291-9_3.

Lipinski, C. A., F. Lombardo, B. W. Dominy und P. J. Feeney (1997). „Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings". In: *Advanced Drug Delivery Reviews* 23.1-3, S. 3–25. DOI: 10.1016/S0169-409X(96)00423-1.

Lolli, G. und R. Battistutta (2013). „Different orientations of low-molecular-weight fragments in the binding pocket of a BRD4 bromodomain". In: *Acta Crystallographica Section D* 69.10, S. 2161–2164. DOI: 10.1107/S090744491301994X.

Lu, H., Y. Xue, G. K. Yu, C. Arias, J. Lin, S. Fong, M. Faure, B. Weisburd, X. Ji, A. Mercier, J. Sutton, K. Luo, Z. Gao und Q. Zhou (2015). „Compensatory induction of MYC expression by sustained CDK9 inhibition via a BRD4-dependent mechanism". In: *eLife* 4. DOI: 10.7554/eLife.06535.

Maggiora, G., M. Vogt, D. Stumpfe und J. Bajorath (2014). „Molecular similarity in medicinal chemistry". In: *Journal of Medicinal Chemistry* 57.8, S. 3186–3204. DOI: 10.1021/jm401411z.

Martin, Y. C., J. L. Kofron und L. M. Traphagen (2002). „Do structurally similar molecules have similar biological activity?" In: *Journal of Medicinal Chemistry* 45.19, S. 4350–4358. DOI: 10.1021/jm020155c.

Meanwell, N. A. (2011). „Synopsis of some recent tactical application of bioisosteres in drug design“. In: *Journal of Medicinal Chemistry* 54.8, S. 2529–2591. DOI: 10.1021/jm1013693.

Meyers, J., N. E. A. Chessum, S. Ali, N. Y. Mok, B. Wilding, A. E. Pasqua, M. Rowlands, M. J. Tucker, L. E. Evans, C. S. Rye, L. O'Fee, Y.-V. Le Bihan, R. Burke, M. Carter, P. Workman, J. Blagg, N. Brown, R. L. M. van Montfort, K. Jones und M. D. Cheeseman (2018). „Privileged structures and polypharmacology within and between protein families“. In: *ACS Medicinal Chemistry Letters* 9.12, S. 1199–1204. DOI: 10.1021/acsmedchemlett.8b00364.

Molecular Operating Environment (MOE), 2013.08 (2013). 1010 Sherbooke St. West, Suite ♯ 910, Montreal, QC, Canada, H3A 2R7.

Moriarty, N. W., R. W. Grosse-Kunstleve und P. D. Adams (2009). „Electronic Ligand builder and optimization workbench (eLBOW): A tool for ligand coordinate and restraint generation“. In: *Acta Crystallographica Section D* 65.Part 10, S. 1074–1080. DOI: 10.1107/S0907444909029436.

Neudert, G. und G. Klebe (2011). „fconv: format conversion, manipulation and feature computation of molecular data“. In: *Bioinformatics* 27.7, S. 1021–1022. DOI: 10.1093/bioinformatics/btr055.

Nikiel, K. (2017). „Effiziente Algorithmen für maximale gemeinsame Graphen zwischen Molekülen unter Berücksichtigung von Bioisosteren“. Masterarbeit. Fakultät für Informatik, TU Dortmund.

O'Boyle, N. M. und R. A. Sayle (2016). „Comparing structural fingerprints using a literature-based similarity benchmark“. In: *Journal of Cheminformatics* 8, S. 36. DOI: 10.1186/s13321-016-0148-0.

Pawar, A., P. N. Gollavilli, S. Wang und I. A. Asangani (2018). „Resistance to BET inhibitor leads to alternative therapeutic vulnerabilities in castration-resistant prostate cancer“. In: *Cell Reports* 22.9, S. 2236–2245. DOI: 10.1016/j.celrep.2018.02.011.

Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng und T. E. Ferrin (2004). „UCSF Chimera–a visualization system for exploratory research and analysis“. In: *Journal of Computational Chemistry* 25.13, S. 1605–1612. DOI: 10.1002/jcc.20084.

Philpott, M., J. Yang, T. Tumber, O. Fedorov, S. Uttarkar, P. Filippakopoulos, S. Picaud, T. Keates, I. Felletar, A. Ciulli, S. Knapp und T. D. Heightman (2011). „Bromodomain-peptide displacement assays for interactome mapping and inhibitor discovery“. In: *Molecular BioSystems* 7.10, S. 2899–2908. DOI: 10.1039/C1MB05099K.

Polvani, S., M. Tarocchi und A. Galli (2012). „PPARγ and oxidative stress: Con(β) catenating NRF2 and FOXO“. In: *PPAR Research* 2012, S. 641087. DOI: 10.1155/2012/641087.

Rathert, P., M. Roth, T. Neumann, F. Muerdter, J.-S. Roe, M. Muhar, S. Deswal, S. Cerny-Reiterer, B. Peter, J. Jude, T. Hoffmann, Boryń, Łukasz M, E. Axelsson, N. Schweifer, U. Tontsch-Grunt, L. E. Dow, D. Gianni, M. Pearson, P. Valent, A. Stark,

N. Kraut, C. R. Vakoc und J. Zuber (2015). „Transcriptional plasticity promotes primary and acquired resistance to BET inhibition“. In: *Nature* 525.7570, S. 543–547. DOI: 10.1038/nature14898.

Raymond, J. W. und P. Willett (2002). „Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases“. In: *Journal of Computer-Aided Molecular Design* 16.1, S. 59–71. DOI: 10.1023/A:1016387816342.

Reddy, A. T., S. P. Lakshmi und R. C. Reddy (2016). „PPARgamma as a novel therapeutic target in lung cancer“. In: *PPAR Research* 2016, S. 8972570. DOI: 10.1155/2016/8972570.

Riniker, S. und G. A. Landrum (2013). „Open-source platform to benchmark fingerprints for ligand-based virtual screening“. In: *Journal of Cheminformatics* 5.1, S. 26. DOI: 10.1186/1758-2946-5-26.

Riveiro, M. E., L. Astorgues-Xerri, R. Vazquez, R. Frapolli, I. Kwee, A. Rinaldi, E. Odore, K. Rezai, M. Bekradda, G. Inghirami, E. Raymond und F. Bertoni (2016). „OTX015 (MK-8628), a novel BET inhibitor, exhibits antitumor activity in non-small cell and small cell lung cancer models harboring different oncogenic mutations“. In: *Oncotarget* 7.51, S. 84675–84687. DOI: 10.18632/oncotarget.13181.

Rogers, D. und M. Hahn (2010). „Extended-connectivity fingerprints“. In: *Journal of Chemical Information and Modeling* 50.5, S. 742–754. DOI: 10.1021/ci100050t.

Rognan, D. und J. Desaphy (2013). „Molecular interaction fingerprints“. In: *Scaffold Hopping in Medicinal Chemistry*. Wiley-Blackwell. Kap. 14, S. 215–230. ISBN: 9783527665143. DOI: 10.1002/9783527665143.ch14.

Ruddigkeit, L., R. van Deursen, L. C. Blum und J.-L. Reymond (2012). „Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17“. In: *Journal of Chemical Information and Modeling* 52.11, S. 2864–2875. DOI: 10.1021/ci300415d.

Sanders, M. P. A., A. J. M. Barbosa, B. Zarzycka, G. A. Nicolaes, J. P. Klomp, J. de Vlieg und A. Del Rio (2012). „Comparative analysis of pharmacophore screening tools“. In: *Journal of Chemical Information and Modeling* 52.6, S. 1607–1620. DOI: 10.1021/ci2005274.

Sauer, W. H. B. und M. K. Schwarz (2003). „Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity“. In: *Journal of Chemical Information and Computer Sciences* 43.3, S. 987–1003. DOI: 10.1021/ci025599w.

Scannell, J. W., A. Blanckley, H. Boldon und B. Warrington (2012). „Diagnosing the decline in pharmaceutical R&D efficiency“. In: *Nature Reviews Drug Discovery* 11, S. 191–200. DOI: 10.1038/nrd3681.

Schäfer, T., N. Kriege, L. Humbeck, K. Klein, O. Koch und P. Mutzel (2017). „Scaffold Hunter: A comprehensive visual analytics framework for drug discovery“. In: *Journal of Cheminformatics* 9.1, S. 28. DOI: 10.1186/s13321-017-0213-3.

Schäfer, T. und P. Mutzel (2017). „StruClus: Scalable structural graph set clustering with representative sampling“. In: *The 13th International Conference on Advanced Data Mining and Applications (ADMA 2017)*. Hrsg. von G. Cong, W.-C. Peng, W. E. Zhang, C. Li und A. Sun. Cham: Springer International Publishing, S. 343–359. ISBN: 978-3-319-69179-4.

Schneider, P. und G. Schneider (2017). „Privileged structures revisited“. In: *Angewandte Chemie International Edition* 56.27, S. 7971–7974. DOI: 10.1002/anie.201702816.

Schuffenhauer, A., P. Ertl, S. Roggo, S. Wetzel, M. A. Koch und H. Waldmann (2007). „The scaffold tree-visualization of the scaffold universe by hierarchical scaffold classification“. In: *Journal of Chemical Information and Modeling* 47.1, S. 47–58. DOI: 10.1021/ci600338x.

Schüttelkopf, A. W. und D. M. F. van Aalten (2004). „PRODRG: A tool for high-throughput crystallography of protein-ligand complexes“. In: *Acta Crystallographica Section D* 60.Part 8, S. 1355–1363. DOI: 10.1107/S0907444904011679.

Shi, J. und C. R. Vakoc (2014). „The mechanisms behind the therapeutic activity of BET bromodomain inhibition“. In: *Molecular Cell* 54.5, S. 728–736. DOI: 10.1016/j.molcel.2014.05.016.

Shimamura, T., Z. Chen, M. Soucheray, J. Carretero, E. Kikuchi, J. H. Tchaicha, Y. Gao, K. A. Cheng, T. J. Cohoon, J. Qi, E. A. Akbay, A. C. Kimmelman, A. L. Kung, J. E. Bradner und K.-K. Wong (2013). „Efficacy of BET bromodomain inhibition in Kras-mutant non-small cell lung cancer“. In: *Clinical Cancer Research*. DOI: 10.1158/1078-0432.CCR-12-3904.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson und D. G. Higgins (2011). „Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega“. In: *Molecular Systems Biology* 7.1. DOI: 10.1038/msb.2011.75.

Smith, S., M. Keul, J. Engel, D. Basu, S. Eppmann und D. Rauh (2017). „Characterization of covalent-reversible EGFR inhibitors“. In: *ACS Omega* 2.4, S. 1563–1575. DOI: 10.1021/acsomega.7b00157.

Song, W.-H., M.-M. Liu, D.-W. Zhong, Y.-l. Zhu, M. Bosscher, L. Zhou, D.-Y. Ye und Z.-H. Yuan (2013). „Tetrazole and triazole as bioisosteres of carboxylic acid: Discovery of diketo tetrazoles and diketo triazoles as anti-HCV agents“. In: *Bioorganic & Medicinal Chemistry Letters* 23.16, S. 4528–4531. DOI: 10.1016/j.bmcl.2013.06.045.

Song, L., Z.-Y. Li, W.-P. Liu und M.-R. Zhao (2015). „Crosstalk between Wnt/β-catenin and Hedgehog/Gli signaling pathways in colon cancer and implications for therapy“. In: *Cancer Biology & Therapy* 16.1, S. 1–7. DOI: 10.4161/15384047.2014.972215.

Sos, M. L., H. B. Rode, S. Heynck, M. Peifer, F. Fischer, S. Klüter, V. G. Pawar, C. Reuter, J. M. Heuckmann, J. Weiss, L. Ruddigkeit, M. Rabiller, M. Koker, J. R. Simard, M. Getlik, Y. Yuza, T.-H. Chen, H. Greulich, R. K. Thomas und D. Rauh (2010). „Chemogenomic profiling provides insights into the limited activity of irreversible

EGFR inhibitors in tumor cells expressing the T790M EGFR resistance mutation“. In: *Cancer Research* 70.3, S. 868–874. DOI: 10.1158/0008-5472.CAN-09-3106.

Sperandio, O., C. H. Reynès, A.-C. Camproux und B. O. Villoutreix (2010). „Rationalizing the chemical space of protein-protein interaction inhibitors“. In: *Drug Discovery Today* 15.5-6, S. 220–229. DOI: 10.1016/j.drudis.2009.11.007.

Stahl, M., H. Mauser, M. Tsui und N. R. Taylor (2005). „A robust clustering method for chemical structures“. In: *Journal of Medicinal Chemistry* 48.13, S. 4358–4366. DOI: 10.1021/jm040213p.

Sterling, T. und J. J. Irwin (2015). „ZINC 15–ligand discovery for everyone“. In: *Journal of Chemical Information and Modeling* 55.11, S. 2324–2337. DOI: 10.1021/acs.jcim.5b00559.

Sturm, N., J. Desaphy, R. J. Quinn, D. Rognan und E. Kellenberger (2012). „Structural insights into the molecular basis of the ligand promiscuity“. In: *Journal of Chemical Information and Modeling* 52.9, S. 2410–2421. DOI: 10.1021/ci300196g.

Tasdemir, N., A. Banito, J.-S. Roe, D. Alonso-Curbelo, M. Camiolo, D. F. Tschaharganeh, C.-H. Huang, O. Aksoy, J. E. Bolden, C.-C. Chen, M. Fennell, V. Thapar, A. Chicas, C. R. Vakoc und S. W. Lowe (2016). „BRD4 connects enhancer remodeling to senescence immune surveillance“. In: *Cancer Discovery* 6.6, S. 612–629. DOI: 10.1158/2159-8290.CD-16-0217.

Taylor, R. D., M. MacCoss und A. D. G. Lawson (2014). „Rings in drugs“. In: *Journal of Medicinal Chemistry* 57.14, S. 5845–5859. DOI: 10.1021/jm4017625.

Terwilliger, T. C. und J. Berendzen (1999). „Evaluation of macromolecular electron-density map quality using the correlation of local r.m.s. density“. In: *Acta Crystallographica Section D* 55.11, S. 1872–1877. DOI: 10.1107/S090744499901029X.

The UniProt Consortium (2017). „UniProt: The universal protein knowledgebase“. In: *Nucleic Acids Research* 45.D1, S. D158–D169. DOI: 10.1093/nar/gkw1099.

Thornber, C. W. (1979). „Isosterism and molecular modification in drug design“. In: *Chemical Society Reviews* 8.4, S. 563–580. DOI: 10.1039/CS9790800563.

Tomassi, S., J. Lategahn, J. Engel, M. Keul, H. L. Tumbrink, J. Ketzer, T. Mühlenberg, M. Baumann, C. Schultz-Fademrecht, S. Bauer und D. Rauh (2017). „Indazole-based covalent inhibitors to target drug-resistant epidermal growth factor receptor“. In: *Journal of Medicinal Chemistry* 60.6, S. 2361–2372. DOI: 10.1021/acs.jmedchem.6b01626.

Varga, T., Z. Czimmerer und L. Nagy (2011). „PPARs are a unique set of fatty acid regulated transcription factors controlling both lipid metabolism and inflammation“. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1812.8. Translating nuclear receptors from health to disease, S. 1007–1022. DOI: 10.1016/j.bbadis.2011.02.014.

Vriend, G. (1990). „WHAT IF: A molecular modeling and drug design program“. In: *Journal of Molecular Graphics* 8.1, S. 52–56. DOI: 10.1016/0263-7855(90)80070-V.

Wallis, W. D., P. Shoubridge, M. Kraetz und D. Ray (2001). „Graph distances using graph union“. In: *Pattern Recognition Letters* 22.6-7, S. 701–704. DOI: 10.1016/S0167-8655(01)00022-8.

Wang, J., W. Wang, P. A. Kollman und D. A. Case (2006). „Automatic atom type and bond type perception in molecular mechanical calculations“. In: *Journal of Molecular Graphics and Modelling* 25.2, S. 247–260. DOI: 10.1016/j.jmgm.2005.12.005.

Wang, J., R. M. Wolf, J. W. Caldwell, P. A. Kollman und D. A. Case (2004). „Development and testing of a general amber force field“. In: *Journal of Computational Chemistry* 25.9, S. 1157–1174. DOI: 10.1002/jcc.20035.

Wang, L., B. Waltenberger, E.-M. Pferschy-Wenzig, M. Blunder, X. Liu, C. Malainer, T. Blazevic, S. Schwaiger, J. M. Rollinger, E. H. Heiss, D. Schuster, B. Kopp, R. Bauer, H. Stuppner, V. M. Dirsch und A. G. Atanasov (2014). „Natural product agonists of peroxisome proliferator-activated receptor gamma (PPARγ): A review“. In: *Biochemical Pharmacology* 92.1, S. 73–89. DOI: 10.1016/j.bcp.2014.07.018.

Wang, Y., S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He und J. Zhang (2017). „PubChem BioAssay: 2017 update“. In: *Nucleic Acids Research* 45.D1, S. D955–D963. DOI: 10.1093/nar/gkw1118.

Waring, M. J., H. Chen, A. A. Rabow, G. Walker, R. Bobby, S. Boiko, R. H. Bradbury, R. Callis, E. Clark, I. Dale, D. L. Daniels, A. Dulak, L. Flavell, G. Holdgate, T. A. Jowitt, A. Kikhney, M. McAlister, J. Méndez, D. Ogg, J. Patel, P. Petteruti, G. R. Robb, M. B. Robers, S. Saif, N. Stratton, D. I. Svergun, W. Wang, D. Whittaker, D. M. Wilson und Y. Yao (2016). „Potent and selective bivalent inhibitors of BET bromodomains“. In: *Nature Chemical Biology* 12, S. 1097–1104. DOI: 10.1038/nchembio.2210.

Welsch, M. E., S. A. Snyder und B. R. Stockwell (2010). „Privileged scaffolds for library design and drug discovery“. In: *Current Opinion in Chemical Biology* 14.3, S. 347–361. DOI: 10.1016/j.cbpa.2010.02.018.

Wermuth, C.-G., C. R. Ganellin, P. Lindberg und L. A. Mitscher (1998). „Glossary of terms used in medicinal chemistry“. In: *Pure and Applied Chemistry* 70.5, S. 1129–1143.

Wetzel, S., K. Klein, S. Renner, D. Rauh, T. I. Oprea, P. Mutzel und H. Waldmann (2009). „Interactive exploration of chemical space with Scaffold Hunter“. In: *Nature Chemical Biology* 5.8, S. 581–583. DOI: 10.1038/nchembio.187.

Wieduwilt, M. J. und M. M. Moasser (2008). „The epidermal growth factor receptor family: Biology driving targeted therapeutics“. In: *Cellular and Molecular Life Sciences* 65.10, S. 1566–1584. DOI: 10.1007/s00018-008-7440-8.

Wilkins, C. L. und M. Randić (1980). „A graph theoretical approach to structure-property and structure-activity correlations“. In: *Theoretica Chimica Acta* 58.1, S. 45–68. DOI: 10.1007/BF00635723.

Winn, M. D., C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin und K. S.

Wilson (2011). „Overview of the CCP4 suite and current developments“. In: *Acta Crystallographica Section D* 67.Part 4, S. 235–242. DOI: 10.1107/S0907444910045749.

Wu, T., G. Wang, W. Chen, Z. Zhu, Y. Liu, Z. Huang, Y. Huang, P. Du, Y. Yang, C.-Y. Liu und L. Cui (2018). „Co-inhibition of BET proteins and NF-κB as a potential therapy for colorectal cancer through synergistic inhibiting MYC and FOXM1 expressions“. In: *Cell Death & Disease* 9.3. DOI: 10.1038/s41419-018-0354-y.

Wyce, A., J. J. Matteo, S. W. Foley, D. J. Felitsky, S. R. Rajapurkar, X.-P. Zhang, M. C. Musso, S. Korenchuk, N. O. Karpinich, K. M. Keenan, M. Stern, L. K. Mathew, C. F. McHugh, M. T. McCabe, P. J. Tummino, R. G. Kruger, C. Carpenter und O. Barbash (2018). „MEK inhibitors overcome resistance to BET inhibition across a number of solid and hematologic cancers“. In: *Oncogenesis* 7.4. DOI: 10.1038/s41389-018-0043-9.

Yin, J.-w. und G. Wang (2014). „The Mediator complex: A master coordinator of transcription and cell lineage development“. In: *Development (Cambridge, England)* 141.5, S. 977–987. DOI: 10.1242/dev.098392.

Zhang, G., A. N. Plotnikov, E. Rusinova, T. Shen, K. Morohashi, J. Joshua, L. Zeng, S. Mujtaba, M. Ohlmeyer und M.-M. Zhou (2013). „Structure-guided design of potent diazobenzene inhibitors for the BET bromodomains“. In: *Journal of Medicinal Chemistry* 56.22, S. 9251–9264. DOI: 10.1021/jm401334s.

Zhang, J., X. Liu, X.-B. Xie, X.-C. Cheng und R.-L. Wang (2016). „Multitargeted bioactive ligands for PPARs discovered in the last decade“. In: *Chemical Biology & Drug Design*. DOI: 10.1111/cbdd.12806.

Zhang, M.-Q. und B. Wilkinson (2007). „Drug discovery beyond the rule-of-five“. In: *Current Opinion in Biotechnology* 18.6, S. 478–488. DOI: 10.1016/j.copbio.2007.10.005.

Internetquellen

Aggregator Advisor (2019). URL: http://advisor.bkslab.org/ (besucht am 21. 01. 2019).

Anchor-Biased MCR Reactions (2018). URL: http://anchorquery.csb.pitt.edu/reactions/ (besucht am 10. 07. 2018).

BLAST - Basic Local Alignment Search Tool (2018). URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi (besucht am 26. 09. 2018).

Cancer Cell Line Encyclopedia (2018). URL: https://portals.broadinstitute.org/ccle (besucht am 07. 08. 2018).

Chemical Probes (2018). URL: http://www.chemicalprobes.org/ (besucht am 16. 07. 2018).

ClinicalTrials (2019). URL: clinicaltrials.gov (besucht am 21. 01. 2019).

Clustal Omega (2018). URL: http://www.clustal.org/omega/ (besucht am 26. 09. 2018).

Clustal Omega - Multiple Sequence Alignment (Webserver) (2018). URL: https://www.ebi.ac.uk/Tools/msa/clustalo/ (besucht am 26. 09. 2018).

DrugBank (2018). URL: https://www.drugbank.ca/ (besucht am 27. 10. 2018).
eMolecules (2018). URL: www.emolecules.com (besucht am 08. 07. 2018).
Epigeneous Bromodomain Assay: BRD4(1) - Cisbio (2018). URL: https://www.cisbio.com/sites/default/files/ressources/cisbio_DD_an_EPIgeneous-Bromodomain-Assay-BRD4%281%29.pdf (besucht am 11. 07. 2018).
ExPASy ProtParam Tool (2018). URL: https://web.expasy.org/protparam/ (besucht am 25. 09. 2018).
FastROCS Toolkit (2018). URL: https://docs.eyesopen.com/toolkits/python/fastrocstk/index.html (besucht am 15. 07. 2018).
fconv atom types (2018). URL: http://pc1664.pharmazie.uni-marburg.de/download/fconv/types (besucht am 27. 09. 2018).
Gnuplot 5.0: An interactive plotting program (2018). URL: http://www.gnuplot.info/ (besucht am 26. 09. 2018).
iPPI-DB Inhibitors of Protein-Protein Interaction Database (2018). URL: http://www.ippidb.cdithem.fr/ (besucht am 08. 07. 2018).
Jarvis-Patrick-Clustering Documentation ChemAxon (2018). URL: https://docs.chemaxon.com/display/docs/Jarvis-Patrick+clustering (besucht am 23. 08. 2018).
MolPort (2018). URL: www.molport.com (besucht am 08. 07. 2018).
RCSB PDB - Research Collaboratory for Structural Bioinformatics Protein Data Bank (2018). URL: rcsb.org (besucht am 09. 07. 2018).
RDKit: Open-source cheminformatics (2018). URL: http://www.rdkit.org (besucht am 26. 09. 2018).
SciFinder (2018). URL: scifinder.cas.org (besucht am 26. 09. 2018).
Theory Toolkits (2018). URL: https://docs.eyesopen.com/toolkits/python/shapetk/shape_theory.html#molecular-shape-section (besucht am 15. 07. 2018).
TR-FRET basics - Cisbio (2018). URL: https://www.cisbio.com/drug-discovery/tr-fret-basics (besucht am 10. 07. 2018).
UniProt (2018). URL: https://www.uniprot.org/ (besucht am 13. 07. 2018).
wwPDB Validation Service (2018). URL: https://validate-rcsb-2.wwpdb.org/ (besucht am 25. 09. 2018).

Anhang

A. Die Atomtypen- und Pharmakophor-Ebene der Ähnlichkeit

A.1. Zuordnung der fconv-Atomtypengruppen zu den fconv-Atomtypen

Index	Atomtyp Kurzform	Atomtyp Langform	Gruppen ID	Gruppen Name
0	Br.0	gebundenes Bromatom	5	Halogen
1	Br.i	Bromidion	6	Halogenid-Ionen
2	C.1n	sp^1-Kohlenstoffatom in Nitrilgruppe		
3	C.1p	sp^1-Kohlenstoffatom mit einem gebun-denen Heavy-Atom*	4	hydrophob
4	C.1s	sp^1-Kohlenstoffatom mit zwei gebun-denen Heavy-Atoms*	4	hydrophob
5	C.2p	anderes sp^2-Kohlenstoffatom mit einem gebundenen Heavy-Atom*		
6	C.2r3	sp^2-Kohlenstoffatom in Dreiring	40	hydrophob, Ring
7	C.2r3o	Carbonylkohlenstoffatom in Cyclopropanon oder Cyclopropenon	3	aromatisch
8	C.2r3x	sp^2-Kohlenstoffatom in heterozyklischem Dreiring	0	Ring
9	C.2s	anderes sp^2-Kohlenstoffatom mit zwei gebundenen Heavy-Atoms*		
10	C.2t	anderes sp^2-Kohlenstoffatom mit drei gebundenen Heavy-Atoms*		
11	C.3n	anderes sp^3-Kohlenstoffatom gebunden an Stickstoffatom		
12	C.3p	anderes sp^3-Kohlenstoffatom mit einem gebundenen Heavy-Atom*	4	hydrophob

Index	Atomtyp Kurzform	Atomtyp Langform	Gruppen ID	Gruppen Name
13	C.3q	anderes sp^3-Kohlenstoffatom mit vier gebundenen Heavy-Atoms*	4	hydrophob
14	C.3r3	sp^3-Kohlenstoffatom in Dreiring	40	hydrophob, Ring
15	C.3r3x	sp^3-Kohlenstoffatom in heterozyklischem Dreiring	0	Ring
16	C.3s	anderes sp^3-Kohlenstoffatom mit zwei gebundenen Heavy-Atoms*	4	hydrophob
17	C.3t	anderes sp^3-Kohlenstoffatom mit drei gebundenen Heavy-Atoms*	4	hydrophob
18	C.am	Carbonylkohlenstoffatom in Amidgruppe		
19	C.ar	sp^2-Kohlenstoffatom in anderem Aromat	3	aromatisch
20	C.ar6	sp^2-Kohlenstoffatom in Benzen	3	aromatisch
21	C.ar6p	sp^2-Kohlenstoffatom mit einer positiv geladenen Resonanzstruktur in einem protonierten heteroaromatischen Sechsring	37	aromatisch, positiv geladen
22	C.ar6x	sp^2-Kohlenstoffatom in einem heteroaromatischen Sechsring	3	aromatisch
23	C.arp	sp^2-Kohlenstoffatom mit einer positiv geladenen Resonanzstruktur in anderen protonierten heteroaromatischen Ringen	37	aromatisch, positiv geladen
24	C.arx	sp^2-Kohlenstoffatom in anderen Heteroaromaten	3	aromatisch
25	C.co2	sp^2-Kohlenstoffatom in COO^--Gruppe (auch gesetzt, wenn der Protonierungszustand unbekannt ist)		
26	C.co2h	sp^2-Kohlenstoffatom in explizit protonierter Carboxygruppe		
27	C.es	Carbonylkohlenstoffatom in Estergruppe oder in Anhydriden		
28	C.et	sp^3-Kohlenstoffatom in Ethern	4	hydrophob
29	C.gu	sp^2-Kohlenstoffatom in unprotonierter Guanidingruppe		
30	C.guh	sp^2-Kohlenstoffatome in protonierter Guanidingruppe (auch gesetzt, wenn der Protonierungszustand unbekannt ist)		
31	C.hal	Carbonylkohlenstoffatom in Säurehalogeniden	5	Halogen

Index	Atomtyp Kurzform	Atomtyp Langform	Gruppen ID	Gruppen Name
32	C.mi	sp^2-Kohlenstoffatom in unprotonierter Amidingruppe		
33	C.mih	sp^2-Kohlenstoffatome in protonierter Amidingruppe (auch gesetzt, wenn der Protonierungszustand unbekannt ist)		
34	C.n	sp^2-Kohlenstoffatom in Iminen		
35	C.o	anderes Carbonylkohlenstoffatom		
36	C.ohp	sp^3-Kohlenstoffatom in primären Alko-holen	124	Doneptor, hydrophob
37	C.ohs	sp^3-Kohlenstoffatom in sekundären Alkoholen	124	Doneptor, hydrophob
38	C.oht	sp^3-Kohlenstoffatom in tertiären Alko-holen	124	Doneptor, hydrophob
39	C.s	Kohlenstoffatom in Sulfinylgruppe		
40	Cl.0	gebundenes Chloratom	5	Halogen
41	Cl.i	Chloridion	6	Halogenid-Ionen
42	F.0	gebundenes Fluoratom	5	Halogen
43	F.i	Fluoridion	6	Halogenid-Ionen
44	H.0	anderes Wasserstoffatom		
45	H.ac	saures Wasserstoffatom (gebunden an O.3ac, N.im, N.sam oder N.ohac)		
46	H.n	Wasserstoffatom gebunden an anderes Stickstoffatom		
47	H.o	Wasserstoffatom gebunden an anderes Sauerstoffatom		
48	H.onh	an ein Stickstoffatom gebundenes Wasserstoffatom in Amidgruppe	12	Doneptor
49	I.0	gebundendes Iodatom	5	Halogen
50	I.i	Iodidion	6	Halogenid-Ionen
51	N.1	anderes sp^1-Stickstoffatom	1	Akzeptor
52	N.2n	sp^2-Stickstoffatom gebunden an anderes Stickstoffatom	1	Akzeptor
53	N.2p	anderes sp^2-Stickstoffatom mit einem gebundenen Heavy-Atom*	12	Doneptor
54	N.2s	anderes sp^2-Stickstoffatom mit zwei gebundenen Heavy-Atoms*	1	Akzeptor

Index	Atomtyp Kurzform	Atomtyp Langform	Gruppen ID	Gruppen Name
55	N.2t	anderes sp^2-Stickstoffatom mit drei gebundenen Heavy-Atoms*		
56	N.3s	sp^3-Stickstoffatom mit zwei gebundenen Heavy-Atoms*	12	Doneptor
57	N.3n	sp^3-Stickstoffatom gebunden an anderes Stickstoffatom	12	Doneptor
58	N.3p	sp^3-Stickstoffatom mit einem gebun-denen Heavy-Atom*	12	Doneptor
59	N.3t	sp^3-Stickstoffatom mit drei gebundenen Heavy-Atoms*	1	Akzeptor
60	N.4q	sp^3-Stickstoffatom mit vier gebundenen Heavy-Atoms*		
61	N.4h	sp^3-Stickstoffatom mit vier gebundenen Atomen (mindestens ein Wasserstoff-atom)	2	Donor
62	N.aap	Stickstoffatom in primären aromatischen Aminen (Hybridisierungszustand kann nicht exakt bestimmt werden)	123	aromatisch, Doneptor
63	N.aas2	sp^2-Stickstoffatom in sekundären aromatischen Aminen	3	aromatisch
64	N.aas3	sp^3-Stickstoffatom in sekundären aromatische Aminen	3	aromatisch
65	N.aat2	sp^2-Stickstoffatom in tertiären aroma-tischen Aminen	3	aromatisch
66	N.aat3	sp^3-Stickstoffatom in tertiären aroma-tischen Aminen	13	aromatisch, Akzeptor
67	N.amp	Stickstoffatom in Carbonsäure- oder Thionamiden mit einem gebundenen Heavy-Atom*	2	Donor
68	N.ams	Stickstoffatom in Carbonsäure- oder Thionamiden mit zwei gebundenen Heavy-Atoms*	2	Donor
69	N.amt	Stickstoffatom in Carbonsäure- oder Thionamiden mit drei gebundenen Heavy-Atoms*		
70	N.ar2	sp^2-Stickstoffatom in Aromat mit zwei gebundenen Atomen	13	aromatisch, Akzeptor
71	N.ar3	sp^2-Stickstoffatom in Aromat mit drei gebundenen Heavy-Atoms*	3	aromatisch

Index	Atomtyp Kurzform	Atomtyp Langform	Gruppen ID	Gruppen Name
72	N.ar3h	sp^2-Stickstoffatom in Aromat mit zwei gebundenen Heavy-Atoms* und einem gebundenen Wasserstoffatom	23	aromatisch, Donor
73	N.ar6	sp^2-Stickstoffatom in aromatischem Sechsring	3	aromatisch
74	N.ar6p	positiv geladenes Stickstoffatom in aromatischem Sechsring	237	aromatisch, Donor, positiv geladen
75	N.arp	sp^2-Stickstoffatom in protoniertem Aromat	23	aromatisch, Donor
76	N.az	mittleres Stickstoffatom in Azidogruppe	7	positiv geladen
77	N.gu1	Imin-Stickstoffatom in unprotonierter Guanidingruppe	12	Doneptor
78	N.gu2	Amin-Stickstoffatom in unprotonierter Guanidingruppe	12	Doneptor
79	N.guh	Stickstoffatom in protonierter Guanidingruppe (auch gesetzt, wenn der Protonierungszustand unbekannt ist)	2	Donor
80	N.ims	Stickstoffatom in Imiden mit zwei gebundenen Heavy-Atoms*	2	Donor
81	N.imt	Stickstoffatom in Imiden mit drei gebundenen Heavy-Atoms*		
82	N.mi1	Imin-Stickstoffatom in unprotonierter Amidingruppe	12	Doneptor
83	N.mi2	Amin-Stickstoffatom in unprotonierter Amidingruppe	12	Doneptor
84	N.mih	Stickstoffatom in protonierter Amidingruppe (auch gesetzt, wenn der Proto-nierungszustand unbekannt ist)	2	Donor
85	N.o2	Stickstoffatom in Nitrogruppe		
86	N.oh	Stickstoffatom in Hydroxylaminen	1	Akzeptor
87	N.ohac	Stickstoffatom in Hydroxamsäuren		
88	N.ohac+-	Stickstoffatom in Hydroxamsäuren	2	Donor
89	N.r3	sp^3-Stickstoffatom in Aziridin- oder Azirin-Ring	120	Doneptor, Ring
90	N.samp	Stickstoffatom in Sulfonamiden mit einem gebundenen Heavy-Atom*	2	Donor

Index	Atomtyp Kurzform	Atomtyp Langform	Gruppen ID	Gruppen Name
91	N.sams	Stickstoffatom in Sulfonamiden mit zwei gebundenen Heavy-Atoms*	2	Donor
92	N.samt	Stickstoffatom in Sulfonamiden mit drei gebundenen Heavy-Atoms*		
93	O.2co2	sp^2-Sauerstoffatom in Carboxygruppe (sp^2-Sauerstoffatom gebunden an C.co2h)	1	Akzeptor
94	O.2es	sp^2-Sauerstoffatom in Estern oder Anhydriden	1	Akzeptor
95	O.2hal	sp^2-Sauerstoffatom in Säurehalogeniden	1	Akzeptor
96	O.2p	sp^2-Sauerstoffatom in OPO_3H^-, PO_3H^- oder POO^-	1	Akzeptor
97	O.2po	sp^2-Sauerstoffatom in P=O (nicht deprotonierte Gruppen)	1	Akzeptor
98	O.2s	sp^2-Sauerstoffatom in OSO_3^-, SO_3^-, POO^- oder deprotonierten Sulfonamiden	1	Akzeptor
99	O.2so	sp^2-Sauerstoffatom in S=O (nicht deprotonierte Gruppen)	1	Akzeptor
100	O.3ac	Sauerstoffatom in Hydroxygruppe in COOH, CSOH, POOHOH, POOH oder SOOOH	12	Doneptor
101	O.3es	sp^3-Sauerstoffatom in Estern oder Anhydriden	1	Akzeptor
102	O.3et	Sauerstoffatom in aliphatischen Ethern	1	Akzeptor
103	O.3eta	Sauerstoffatom in aromatischen Ethern	13	Akzeptor, aromatisch
104	O.3oh	Sauerstoffatom in Hydroxygruppe	12	Doneptor
105	O.3po	sp^3-Sauerstoffatom mit zwei gebundenen Heavy-Atoms* davon mindestens ein Phosphoratom	1	Akzeptor
106	O.3so	sp^3-Sauerstoffatom mit zwei gebundenen Heavy-Atoms* davon mindestens ein Schwefelatom	1	Akzeptor
107	O.am	Sauerstoffatom in Carbonsäureamiden	1	Akzeptor
108	O.ar	aromatisches Sauerstoffatom	3	aromatisch
109	O.carb	Sauerstoffatom in anderen Carbonylgruppen	1	Akzeptor
110	O.co2	Sauerstoffatom in COO^- oder CSO^-	1	Akzeptor
111	O.h2o	Sauerstoffatom in Wasser	12	Doneptor
112	O.n	Sauerstoffatom in Nitrogruppe	1	Akzeptor

Index	Atomtyp Kurzform	Atomtyp Langform	Gruppen ID	Gruppen Name
113	O.noh	sp^3-Sauerstoffatom in Hydroxylaminen oder Hydroxamsäuren	12	Doneptor
114	O.o	Sauerstoffatom in Peroxidgruppe	1	Akzeptor
115	O.ph	Sauerstoffatom in phenolischer Hydroxygruppe	12	Doneptor
116	O.r3	Sauerstoffatom in Oxiran	10	Akzeptor, Ring
117	P.3	anderes sp^3-Phosphoratom	1	Akzeptor
118	P.o	Phosphoratom in PO		
119	P.o2	Phosphoratom in deprotonierter PO_2-Gruppe (oder mit unbekanntem Protonierungszustand)		
120	P.o2h	Phosphoratom in nicht deprotonierter PO_2-Gruppe		
121	P.o3	Phosphoratom in deprotonierter PO_3-Gruppe (oder mit unbekanntem Protonierungszustand)		
122	P.o3h	Phosphoratom in nicht deprotonierter PO_3-Gruppe		
123	P.o4	Phosphoratom in deprotonierter PO_4-Gruppe (oder mit unbekanntem Protonierungszustand)		
124	P.o4h	Phosphoratom in nicht deprotonierter PO_4-Gruppe		
125	P.r3	Phosphoratom in Phosphiran	0	Ring
126	S.2	Schwefelatom in CSO^-, COS^- oder anderes sp^2-Schwefelatom	1	Akzeptor
127	S.3	anderes sp^3-Schwefelatom	1	Akzeptor
128	S.ar	aromatisches Schwefelatom	3	aromatisch
129	S.o	Schwefelatom in SO		
130	S.o2	Schwefelatom in deprotonierten Sulfonamiden (oder mit unbekanntem Protonierungszustand)		
131	S.o2h	Schwefelatom in protonierten Sulfonamiden oder anderes (nicht S.o2) Schwefelatom in SO_2		
132	S.o3	Schwefelatom in SO_3^- (oder mit unbekanntem Protonierungszustand)		

Index	Atomtyp Kurzform	Atomtyp Langform	Gruppen ID	Gruppen Name
133	S.o3h	Schwefelatom in SO_3		
134	S.o4	Schwefelatom in OSO_3^- (oder mit unbekanntem Protonierungszustand)		
135	S.o4h	Schwefelatom in $OSO_3$3		
136	S.r3	Schwefelatom in Thiiran	10	Akzeptor, Ring
137	S.s	Schwefelatom in Disulfidbindung	1	Akzeptor
138	S.sh	Schwefelatom in Thiolgruppe	12	Doneptor
139	S.thi	Schwefelatom in Sulfinylgruppe	1	Akzeptor

Tabelle A.1.: Zuordnung der fconv-Atomtypengruppen zu den fconv-Atomtypen verändert nach Neudert und Klebe (2011) und *fconv atom types* (2018). Ist keine Atomtypengruppe dem fconv-Atomtyp zugeordnet, so entspricht das Label des Atoms seinem fconv-Atomtyp. Heavy-Atom*: Atom, welches nicht ein Wasserstoffatom ist.

A.2. Verteilung der Scores von ECFP4-TC-Ansätzen und MSS-Ansatz

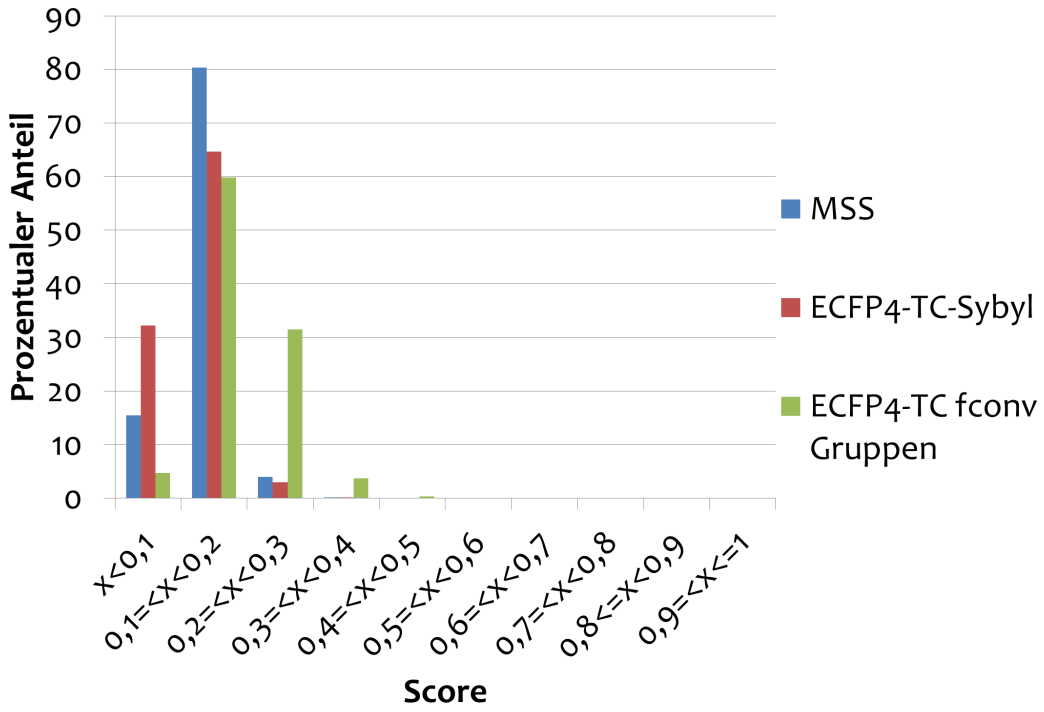


Abbildung A.1.: Score-Verteilung des ECFP4-TC-Ansatzes mit Sybyl-Atomtypen (rot) oder mit fconv-Atomtypengruppen (grün) sowie des MSS-Ansatzes (blau) für die Ähnlichkeitssuche mit Atenolol als Referenz in der ChEMBL.

A.3. Korrelation der Scores von ECFP4-TC-Ansätzen und MSS-Ansatz

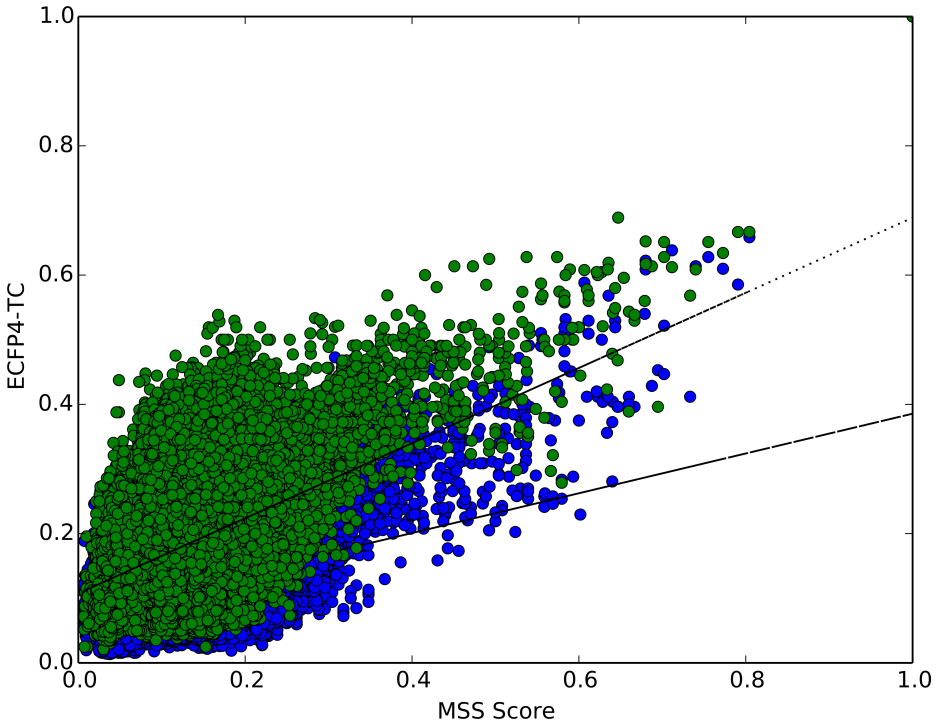
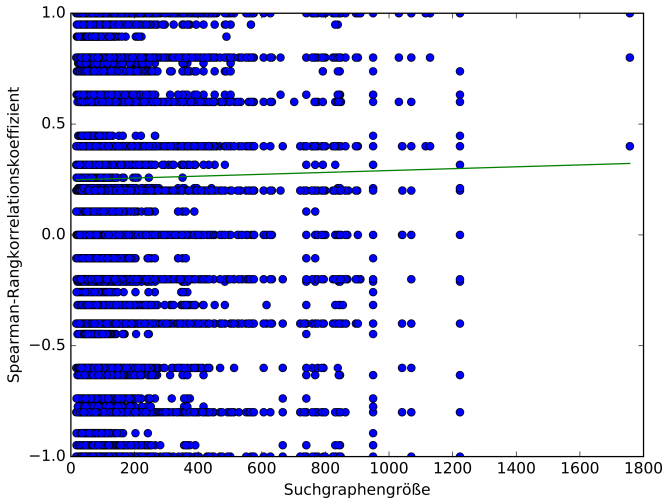


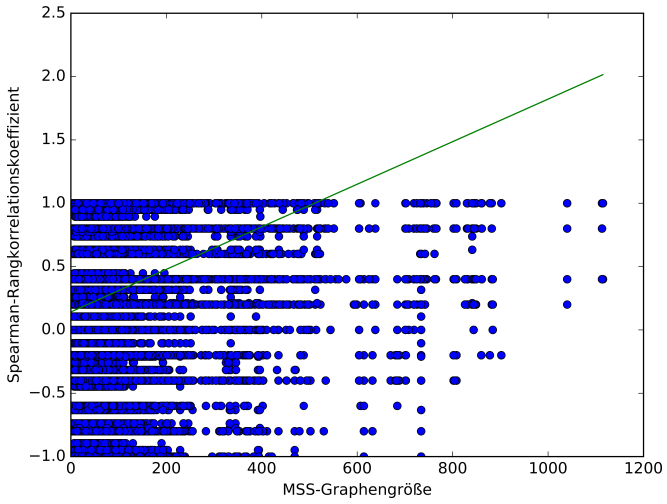
Abbildung A.2.: Korrelation der Scores des ECFP4-TC-Ansatzes mit Sybyl-Atomtypen mit den MSS-Scores (blaue Punkte) sowie Korrelation der Scores des ECFP4-TC-Ansatzes mit fconv-Atomtypengruppen mit den MSS-Scores (grüne Punkte) für die Ähnlichkeitssuche mit Atenolol als Referenz in der ChEMBL. Auf der Abszisse ist der Score nach dem MSS-Ansatz und auf der Ordinate der Score nach ECFP4-TC aufgetragen. Schwarz eingezeichnet ist die berechnete Korrelationsgerade unter Annahme einer linearen Korrelation.

Die Korrelationsgeraden in Abbildung A.2 erscheinen optisch mit zu geringer Steigung im Vergleich zu den Datenpunkten. Dies liegt daran, dass deutlich mehr Daten für den niedrigen Score-Bereich vorliegen, als für den hohen Score-Bereich. Dadurch können nicht alle Datenpunkte im niedrigen Score-Bereich einzeln wahrgenommen und dargestellt werden, was zu einer Verzerrung der intuitiven Wahrnehmung führt. Bemerkenswert ist, dass beide Korrelationsgeraden nicht dem Koordinatenursprung entspringen. Es kann geschlussfolgert werden, dass es beim ECFP4-TC zu einem Grundrauschen durch eine Basisähnlichkeit kommt, welche beim MSS-Ansatz stark reduziert ist. Ähnliche Erkenntnisse wurden bereits von Godden, Xue und Bajorath (2000) gewonnen, sie stellten eine hohe Präferenz des Tanimoto-Koeffizienten für Werte um 1/3 fest.

A.4. Korrelation des Spearman-Rangkorrelationskoeffizienten mit der Graphgröße



(a) Eingabegraphgröße



(b) MSS-Größe

Abbildung A.3.: Korrelation des Spearman-Rangkorrelationskoeffizienten mit der Graphgröße des Eingabegraphen (a) und der MSS-Größe (b) für den Ansatz der Pharmakophor-Ebene unter Beibehaltung der Wasserstoffatome („ph4_H“) auf dem OBDS. Die MSS-Größe wurde für den MSS des Suchgraphen mit dem ersten, d. h. ähnlichsten, Graphen ermittelt. Auf der Abszisse ist die Graphgröße und auf der Ordinate der Spearman-Rangkorrelationskoeffizient aufgetragen. Schwarz eingezeichnet ist die berechnete Korrelationsgerade unter Annahme einer linearen Korrelation.

A.5. Abkürzungen der Fingerprints des OBDS

Abkürzung	Beschreibung	Fingerprint-Klasse
AP	Atompaar Fingerprint (*Count*)	Hash-basiert
Avalon	entwickelt zur Substrukturfilterung bei Suchen	Fragment-basiert
ECFC0	*Count* Form von ECFP0	Hash-basiert
ECFC2	*Count* Form von ECFP2	Hash-basiert
ECFC4	*Count* Form von ECFP4	Hash-basiert
ECFC6	*Count* Form von ECFP6	Hash-basiert
ECFP0	Extended-connectivity-Fingerprint mit Durchmesser 0	Hash-basiert
ECFP2	Extended-connectivity-Fingerprint mit Durchmesser 2	Hash-basiert
ECFP4	Extended-connectivity-Fingerprint mit Durchmesser 4	Hash-basiert
ECFP6	Extended-connectivity-Fingerprint mit Durchmesser 6	Hash-basiert
FCFC2	*Count* Form von FCFP2	Hash-basiert
FCFC4	*Count* Form von FCFP4	Hash-basiert
FCFC6	*Count* Form von FCFP6	Hash-basiert
FCFP2	Feature-class-Fingerprint mit Durchmesser 2	Hash-basiert
FCFP4	Feature-class-Fingerprint mit Durchmesser 4	Hash-basiert
FCFP6	Feature-class-Fingerprint mit Durchmesser 6	Hash-basiert
HashAP	Bitvektor Form von AP	Hash-basiert
HashTT	Bitvektor Form von TT	Hash-basiert
LAvalon	16384 bit Form von Avalon	Fragment-basiert
LECFP4	16384 bit Form von ECFP4	Hash-basiert
LECFP6	16384 bit Form von ECFP6	Hash-basiert
LFCFP4	16384 bit Form von FCFP4	Hash-basiert
LFCFP6	16384 bit Form von FCFP6	Hash-basiert
MACCS	Molecular ACCess System structural keys	Fragment-basiert
RDK5	kodiert Pfade mit maximaler Länge 5	Hash-basiert
RDK6	kodiert Pfade mit maximaler Länge 6	Hash-basiert
RDK7	kodiert Pfade mit maximaler Länge 7	Hash-basiert
TT	Topological torsion Fingerprint (*Count*)	Hash-basiert

Tabelle A.2.: Übersicht der Fingerprints des OBDS verändert nach O'Boyle und Sayle (2016). *Count*: Die Anzahl wie oft ein Merkmal vorkommt, wird gespeichert.

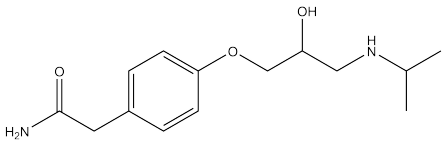
A.6 Vergleich der Ergebnisse der Proof of Concept Studie

Rang MSS	Rang MSS fortl.	ChEMBL-ID	Target UniProt-ID Organismus	Rang ECFP4-TC Sybyl	Rang ECFP4-TC Sybyl fortl.	Rang ECFP4-TC fconv	Rang ECFP4-TC fconv fortl.
1	1	Atenolol	P08588	1	1	1	1
2	2	CHEMBL143467	P18090 (Ratte/Rattus norvegicus)	2	2	3	3
3	3	CHEMBL7154	P18090	11	12	3	4
4	4	Prenalterol (CHEMBL1160714)	P08588	7	8	6	8
5	5	CHEMBL3247307	P18090 (Ratte/Rattus norvegicus)	4	4	5	6
6	6 und 7	Metoprolol (CHEMBL13 bzw. CHEMBL1256957)	P08588	6	6 und 7	13	20
7	8	CHEMBL1159888	Q28044* (Kuh/Bos taurus)	45	65	26	40
8	9	Esmolol (CHEMBL768)	P08588	3	3	12	18
9	10	Toliprolol (CHEMBL67096)	P07550* B0FL73 (Meerschwein/Cavia porcellus)	17	20	5	7
9	11	CHEMBL6920	Q28044* (Kuh/Bos taurus)	32	45	7	11
10	12	CHEMBL3228930	B0FL73 (Meerschwein/Cavia porcellus)	30	41	166	894
11	13	CHEMBL1159891	Q28044* (Kuh/Bos taurus)	37	51	11	15
12	14	CHEMBL3247301	P18090 (Ratte/Rattus norvegicus)	5	5	10	14
12	15	Practolol (CHEMBL6995)	P08588	8	9	4	5

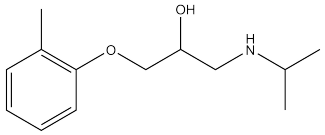
Rang MSS	Rang MSS fortl.	ChEMBL-ID	Target UniProt-ID Organismus	Rang ECFP4-TC Sybyl	Rang ECFP4-TC Sybyl fortl.	Rang ECFP4-TC fconv	Rang ECFP4-TC fconv fortl.
13	16	CHEMBL6939	P18090 (Ratte/Rattus norvegicus) B0FL73 (Meerschwein/Cavia porcellus)	13	14	28	43
14	17	CHEMBL26917	P08588	54	95	40	59
14	18	CHEMBL6902	P18090 (Ratte/Rattus norvegicus)	14	15	35	53
15	19	CHEMBL1159890	Q28044* (Kuh/Bos taurus)	45	66	33	49
16	20	CHEMBL3228931	B0FL73 (Meerschwein/Cavia porcellus)	51	82	178	1046
17	21	CHEMBL31608	B0FL73 (Meerschwein/Cavia porcellus)	9	10	17	28

Tabelle A.3.: Ergebnisse der Proof of Concept Studie zum Vergleich von ECFP4-TC (State of the Art) mit Sybyl-Atomtypen und mit fconv-Atomtypengruppen gegen den MSS-Ansatz. Sofern es sich bei dem Target um ein nicht humanes Protein handelt, wurde der Organismus in Klammern angegeben. * β_2-Adrenozeptor.

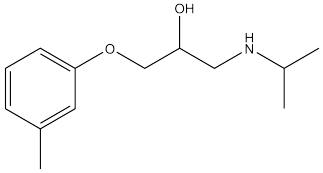
A.7. Analyse der diskriminativen Kraft an einem Beispiel aus der Proof of Concept Studie



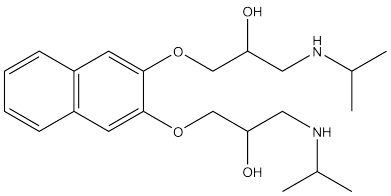
(a) Atenolol



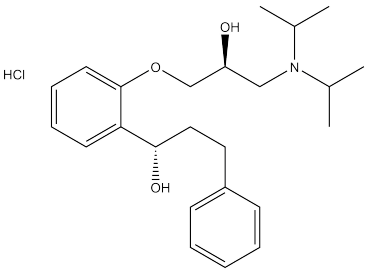
(b) CHEMBL6920
ECFP4-TC: ca. 0,446809
MSS: 0,702128



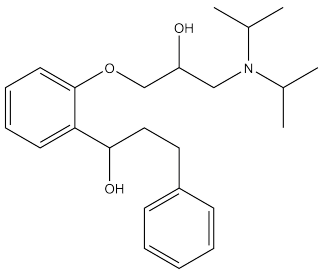
(c) CHEMBL67096
ECFP4-TC: ca. 0,521739
MSS: 0,702128



(d) CHEMBL3272440
ECFP4-TC: ca. 0,446809
MSS: 0,42953



(e) CHEMBL2449367
ECFP4-TC: ca. 0,238095
MSS: 0,42953

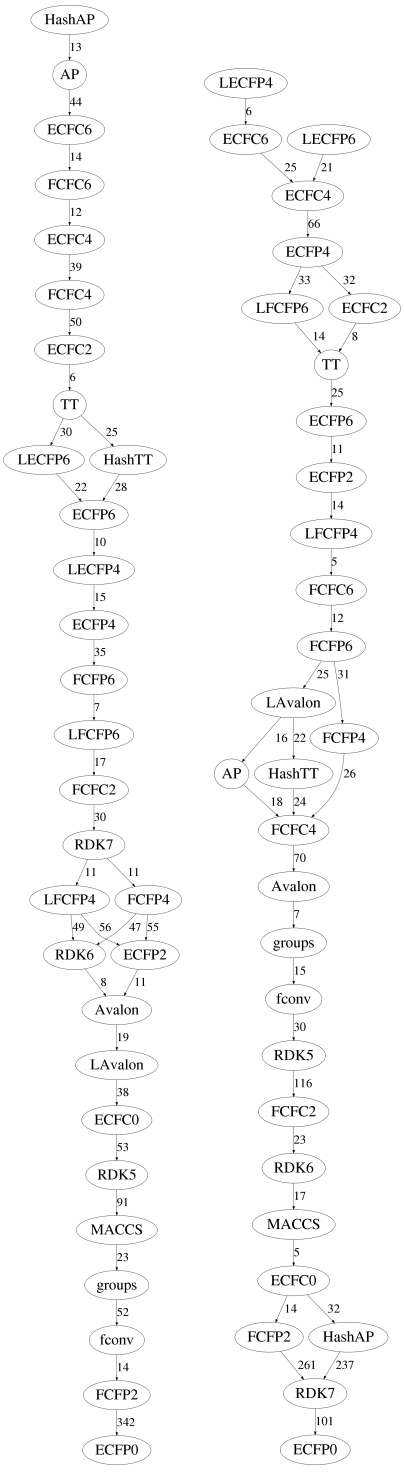


(f) CHEMBL356974
ECFP4-TC: ca. 0,238095
MSS: 0,42953

Abbildung A.4.: Strukturen der Hits der MSS-Suche nach Atenolol (a), welche nicht durch den ECFP4-TC unterschieden werden können: CHEMBL6920 (b) und CHEMBL3272440 (d) sowie die dazugehörigen Strukturen, welche nicht per MSS-Ansatz unterschieden werden können: CHEMBL67096 (c), CHEMBL2449367 (e) und CHEMBL356974 (f).

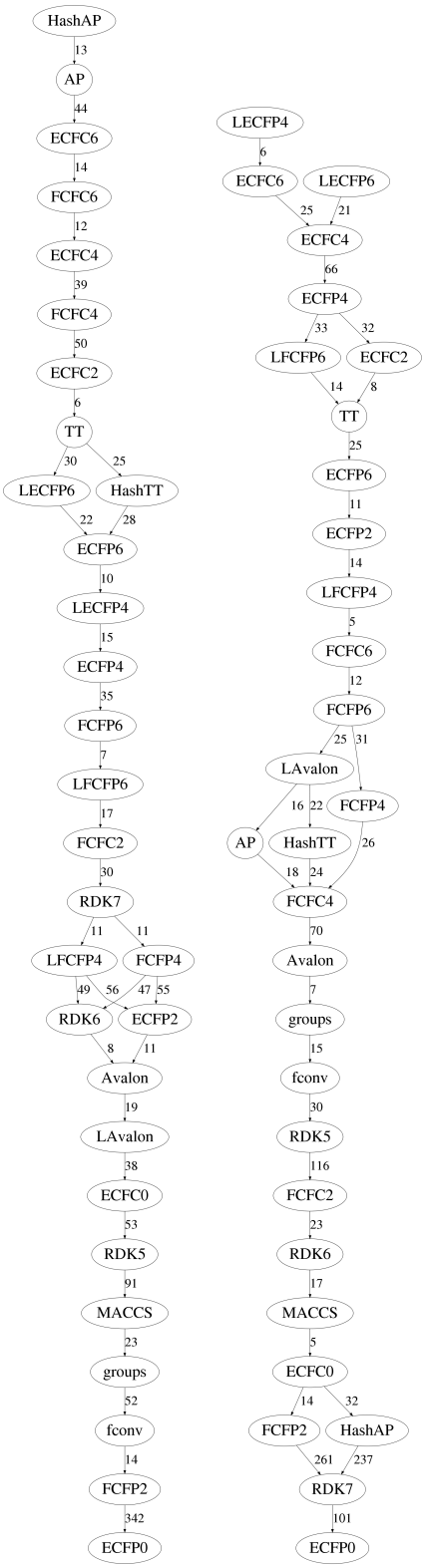
A.8. Ergebnisse der MSS-Suchen für den OBDS

A.8.1. fconv-Atomtypen- und fconv-Atomtypengruppen-Ebene inklusive Test des ct-Parameters



(a) Singel-Assay (b) Multi-Assay

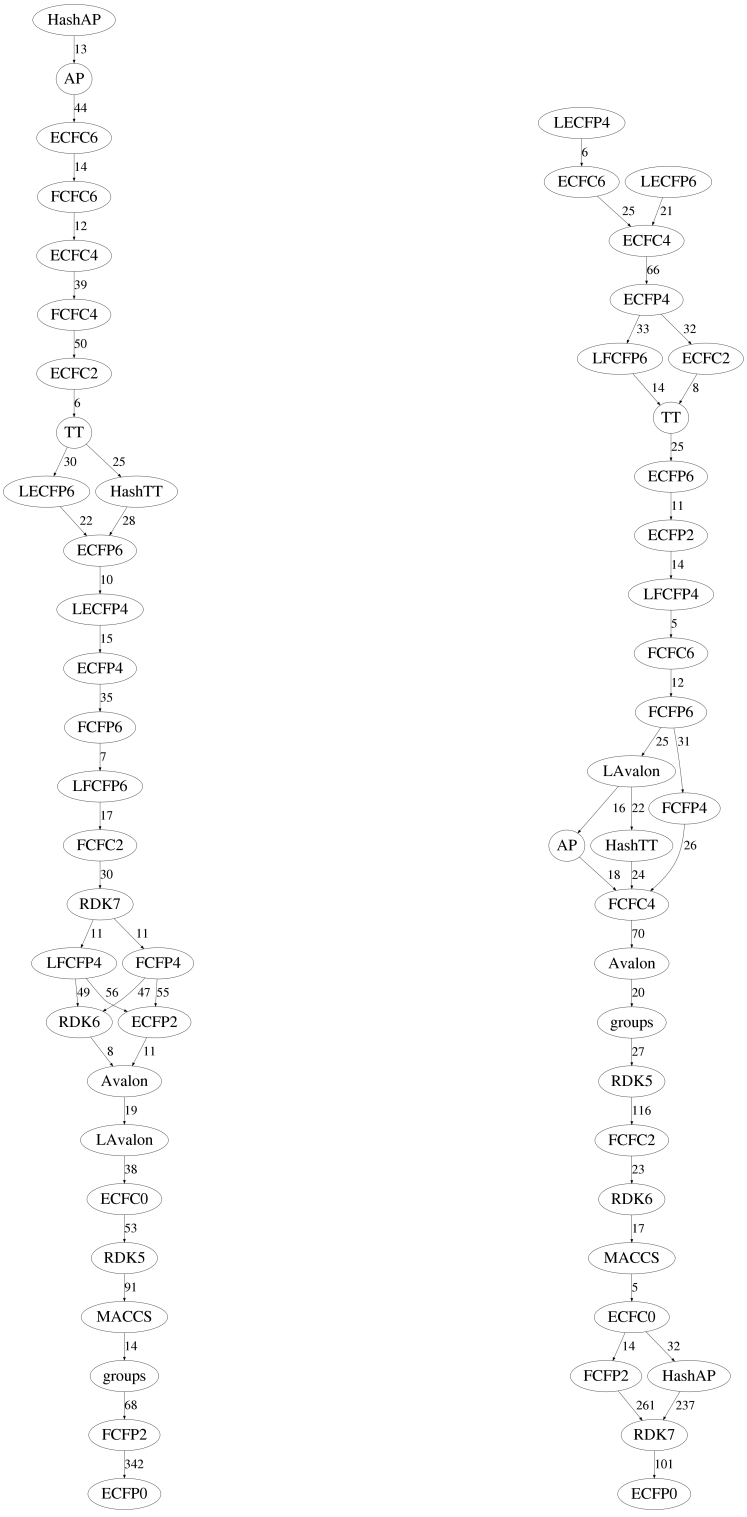
Abbildung A.5.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypen- und Atomtypengruppen-Ebene (Knotenlabel „fconv“ beziehungsweise „groups“) mit Zeitbegrenzung von 1000 Millisekunden (ms).



(a) Singel-Assay (b) Multi-Assay

Abbildung A.6.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypen- und Atomtypengruppen-Ebene (Knotenlabel „fconv“ beziehungsweise „groups“) mit Zeitbegrenzung von 2000 ms.

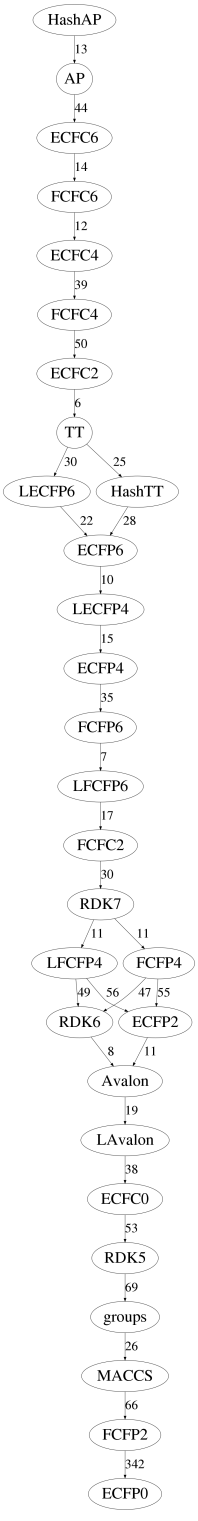
A.8.2. fconv-Atomtypengruppen-Ebene inklusive Ähnlichkeitsmatrix 1



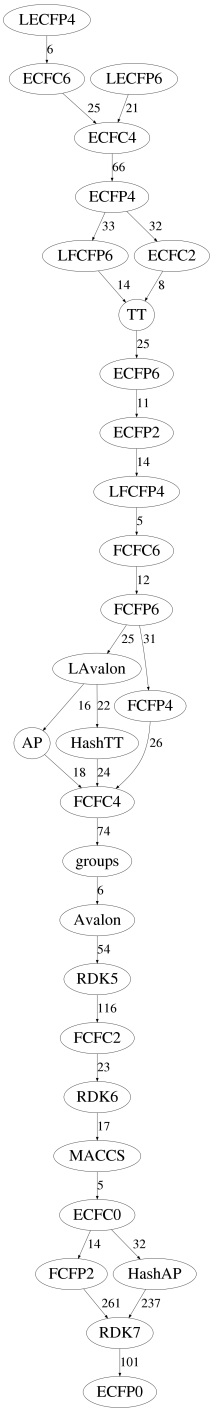
(a) Singel-Assay (b) Multi-Assay

Abbildung A.7.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypengruppen-Ebene und Ähnlichkeitsmatrix 1 (Knotenlabel „groups“).

A.8.3. fconv-Atomtypengruppen-Ebene inklusive Ähnlichkeitsmatrix 2



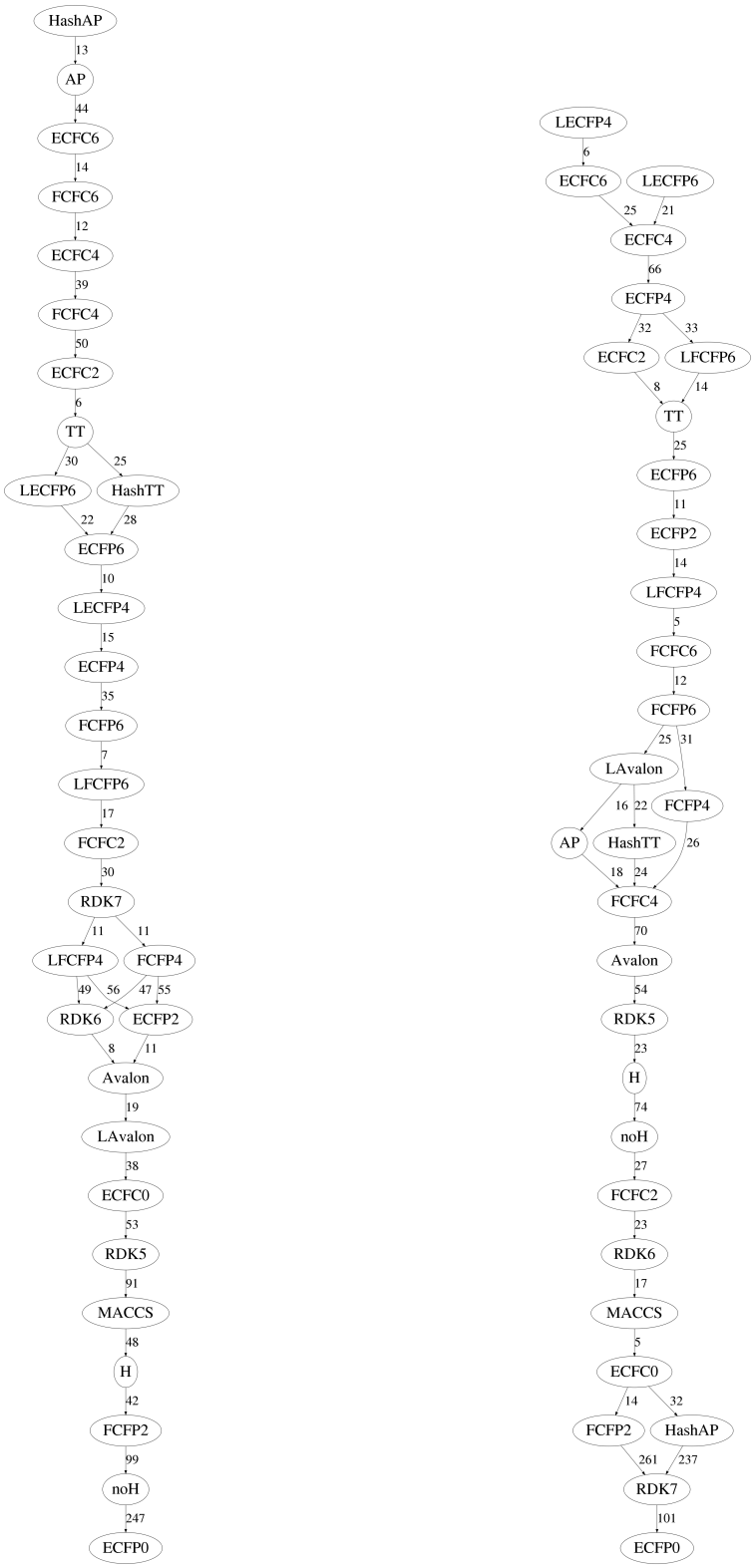
(a) Singel-Assay



(b) Multi-Assay

Abbildung A.8.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Atomtypengruppen-Ebene und Ähnlichkeitsmatrix 2 (Knotenlabel „groups“).

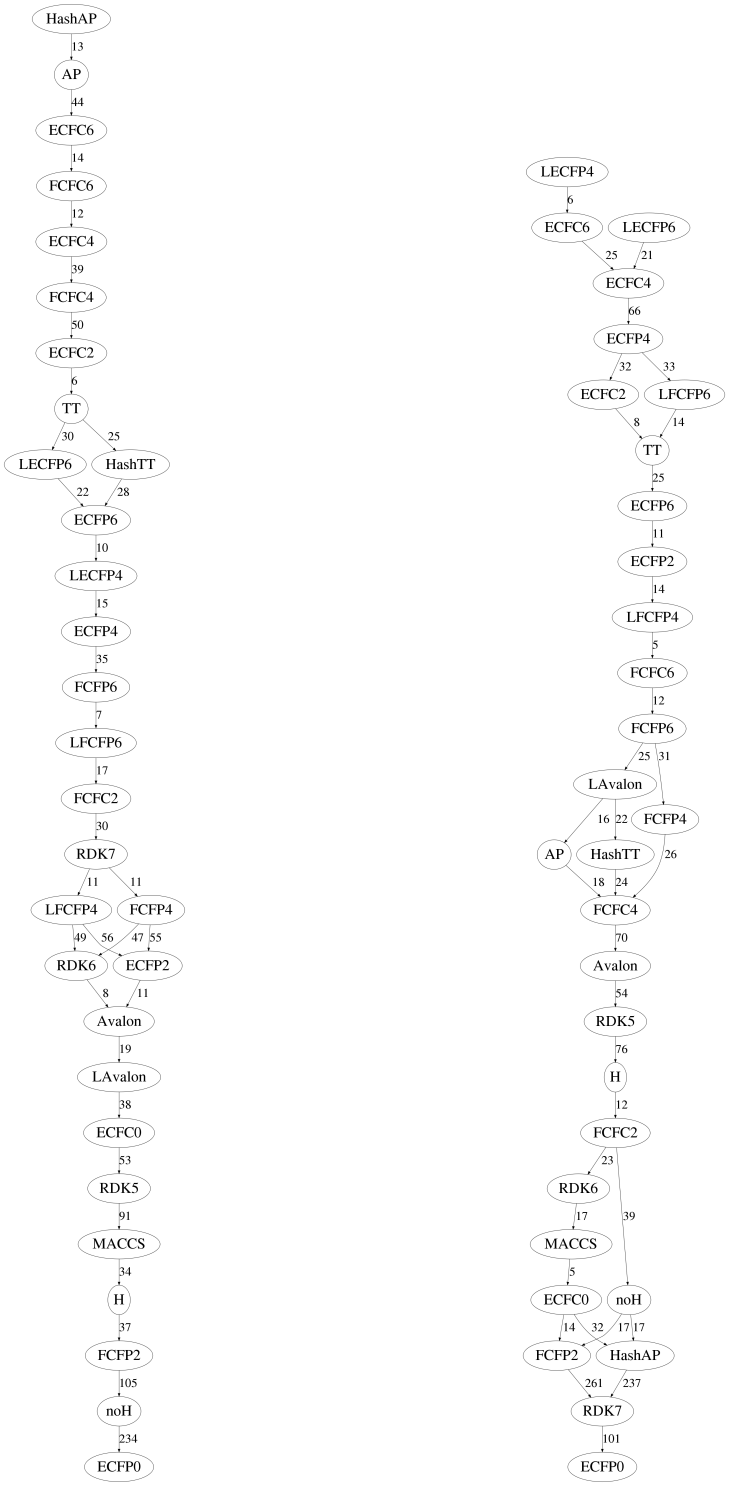
A.8.4. Pharmakophor-Ebene



(a) Singel-Assay (b) Multi-Assay

Abbildung A.9.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Pharmakophor-Ebene (Knotenlabel „H“ unter Berücksichtigung von Wasserstoffatomen und „noH“ ohne Berücksichtigung von Wasserstoffatomen).

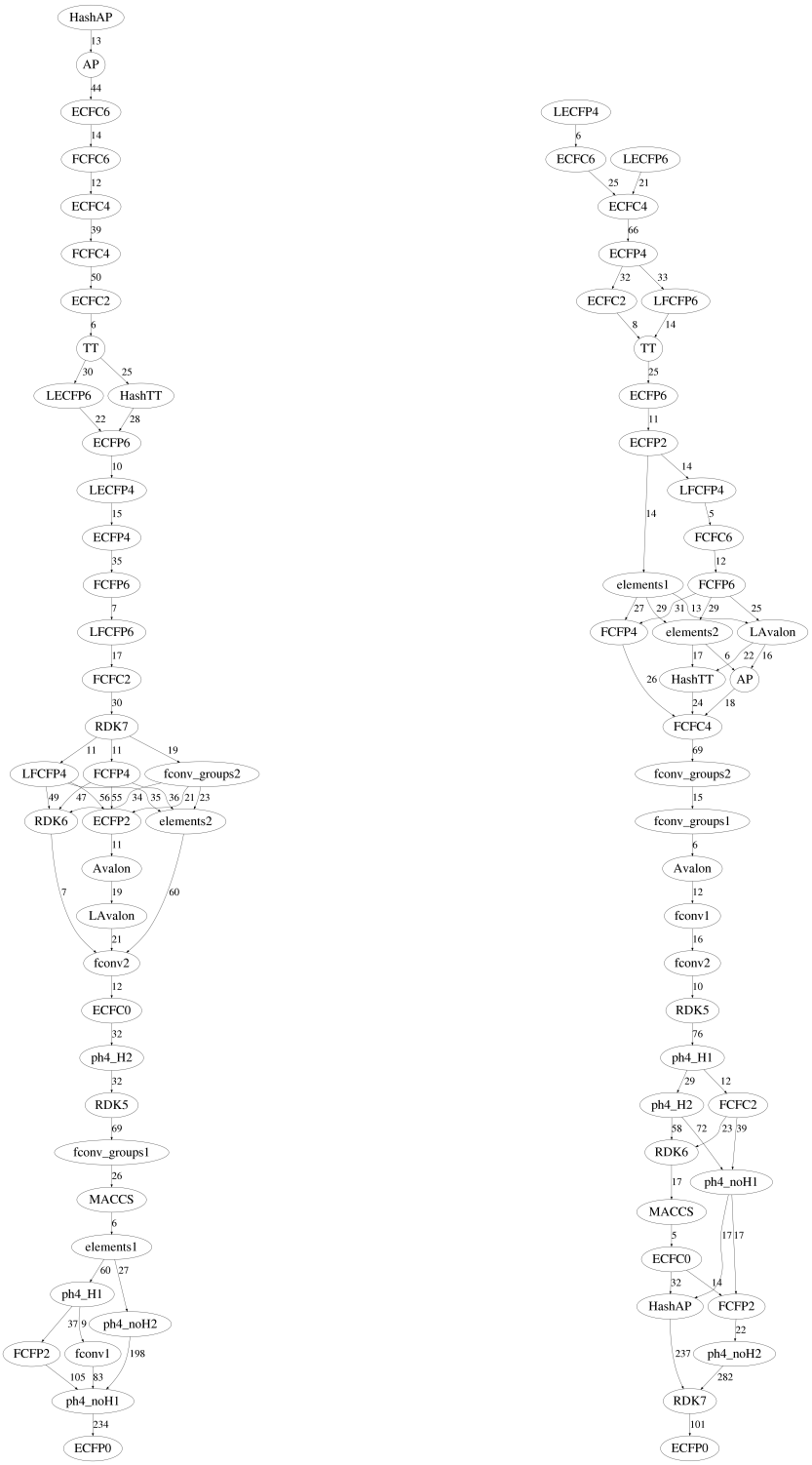
A.8.5. Pharmakophor-Ebene inklusive Ähnlichkeitsmatrix 2



(a) Singel-Assay (b) Multi-Assay

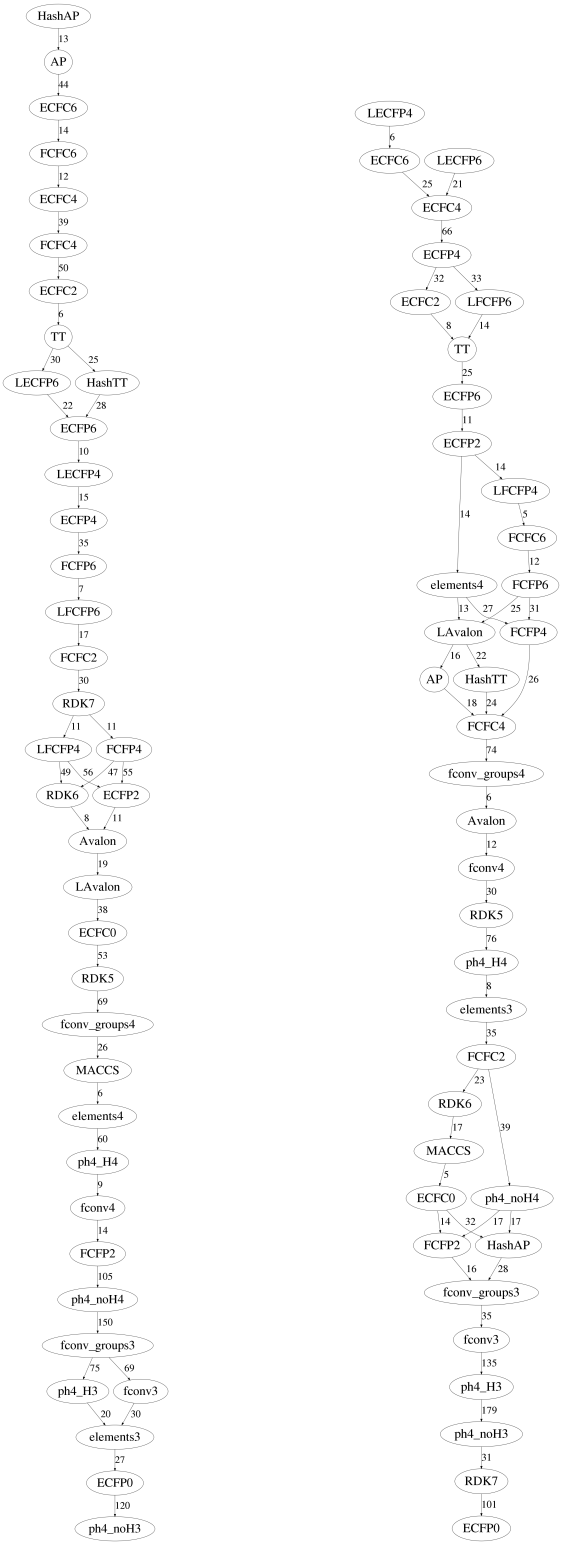
Abbildung A.10.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS basierend auf der Pharmakophor-Ebene und Ähnlichkeitsmatrix 2 (Knotenlabel „H“ unter Berücksichtigung von Wasserstoffatomen und „noH“ ohne Berücksichtigung von Wasserstoffatomen).

A.8.6. Analyse verschiedener Ähnlichkeitsfunktionen



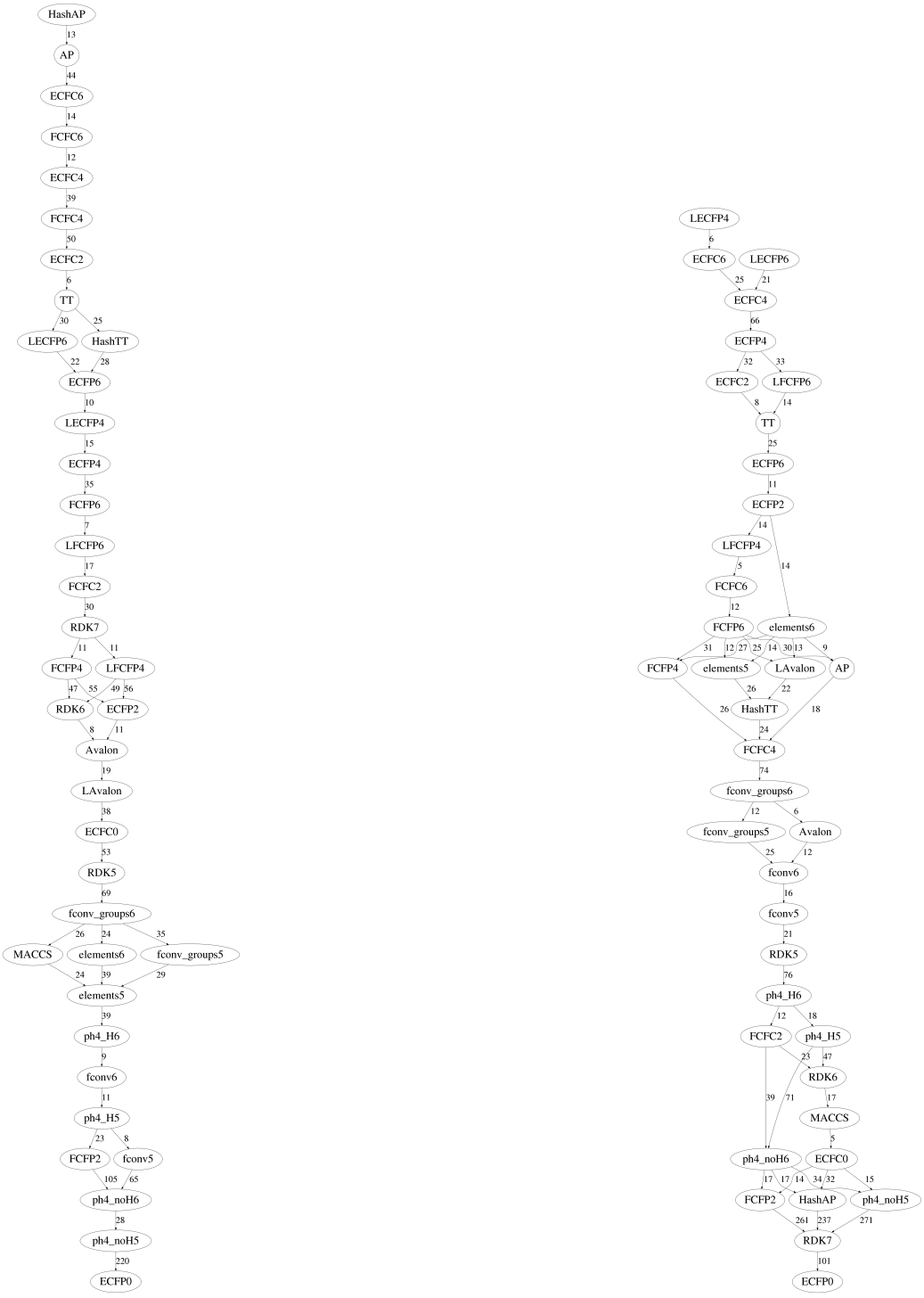
(a) Singel-Assay 1 und 2 (b) Multi-Assay 1 und 2

Abbildung A.11.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 1 und 2 (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen, Zahl: Ähnlichkeitsfunktion).



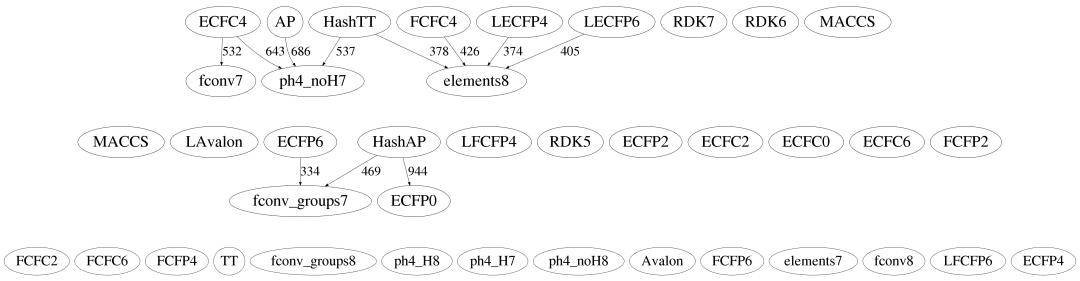
(a) Singel-Assay 3 und 4 (b) Multi-Assay 3 und 4

Abbildung A.12.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 3 und 4 (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen, Zahl: Ähnlichkeitsfunktion).

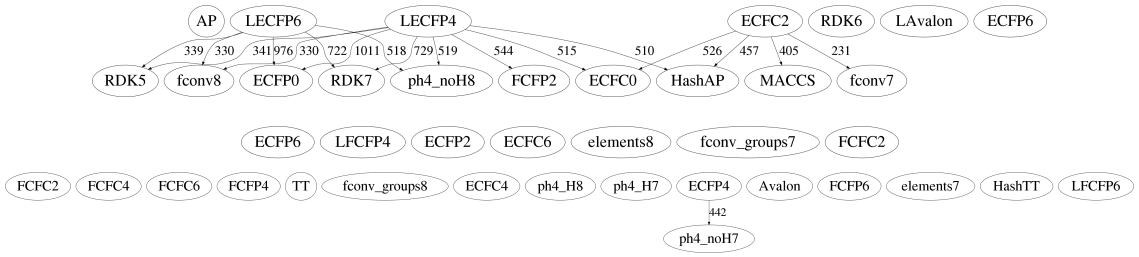


(a) Singel-Assay 5 und 6 (b) Multi-Assay 5 und 6

Abbildung A.13.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 5 und 6 (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen, Zahl: Ähnlichkeitsfunktion).



(a) Singel-Assay 7 und 8



(b) Multi-Assay 7 und 8

Abbildung A.14.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Verwendung der Ähnlichkeitsfunktionen 7 und 8 (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen, Zahl: Ähnlichkeitsfunktion).

A.8.7. Zulassung von Lücken

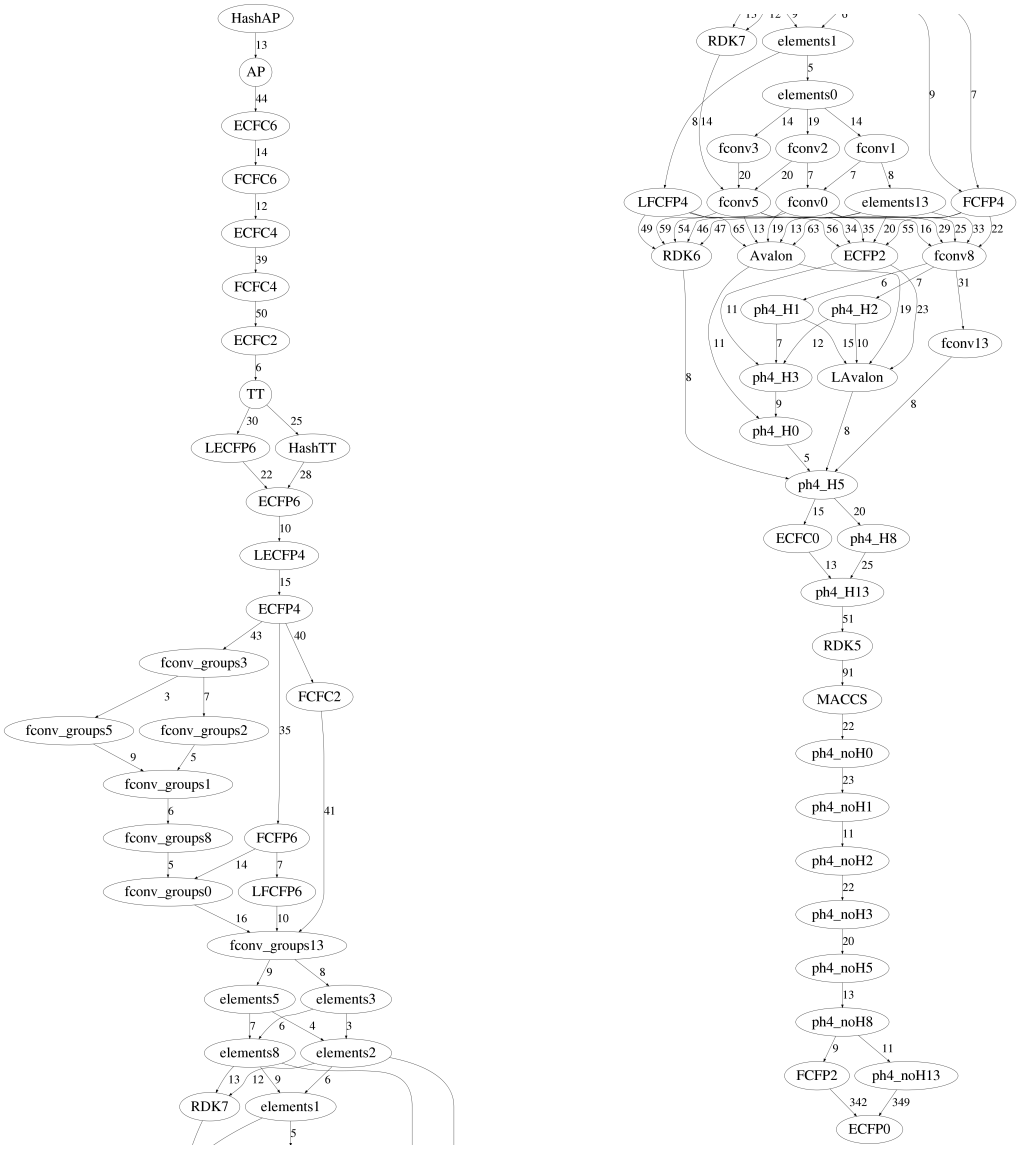


Abbildung A.15.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Zulassung von Lücken für den Single-Assay Benchmarkdatensatz (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen, Zahl: Gap-Strafe).

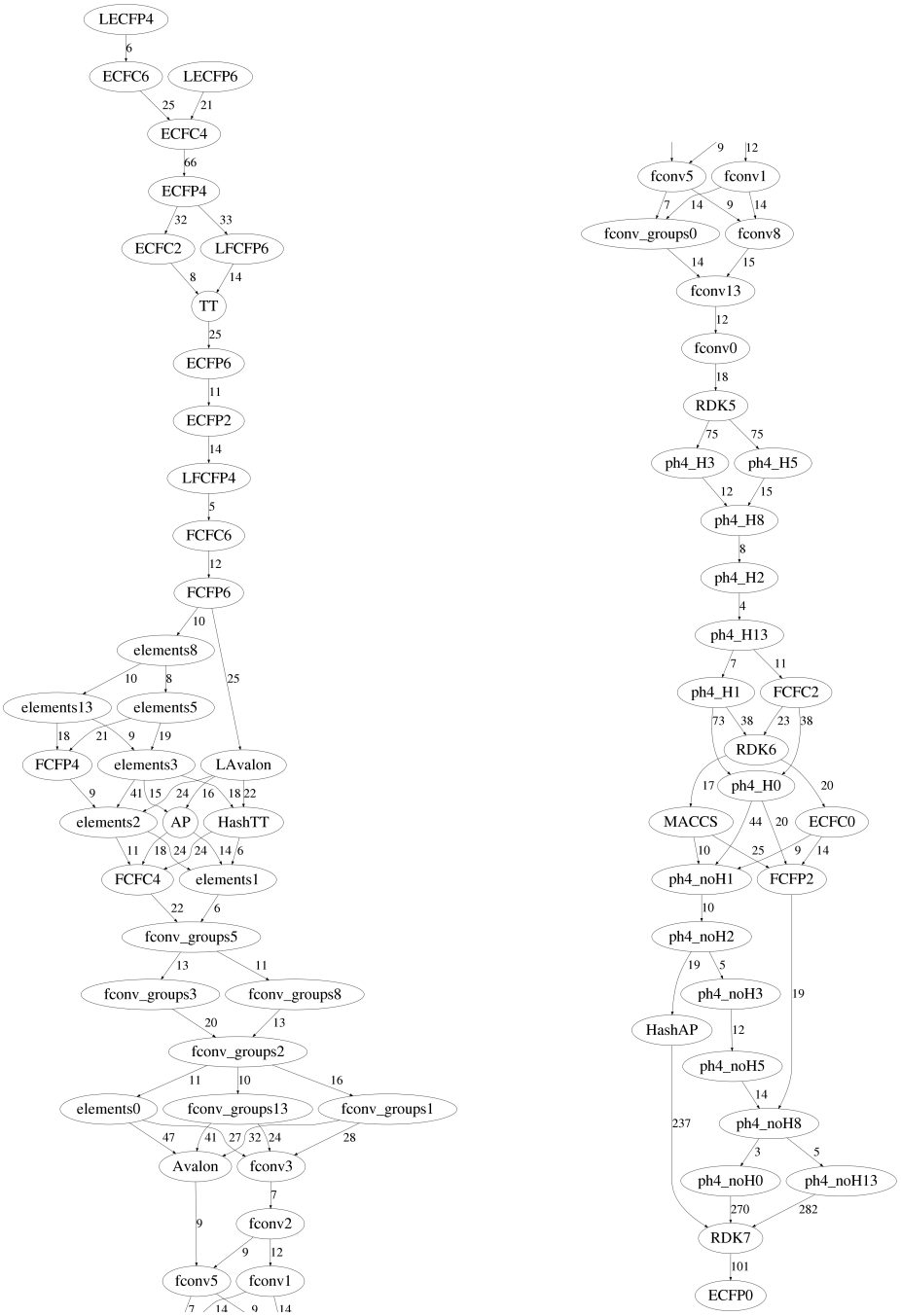
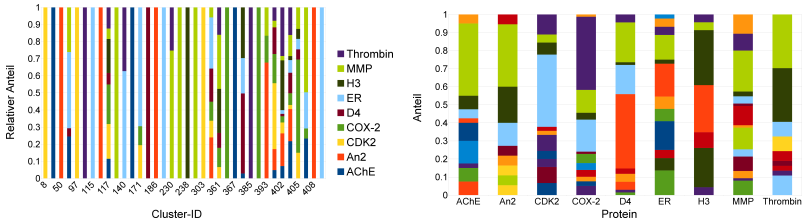


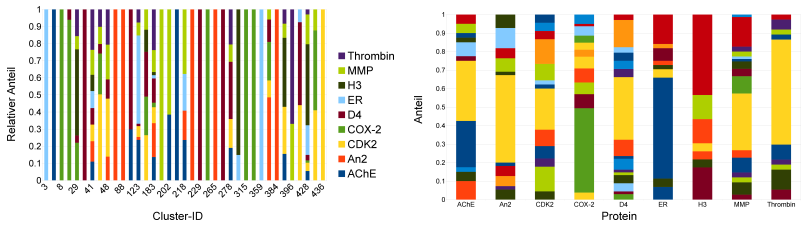
Abbildung A.16.: Performanz einer Ähnlichkeitssuche mittels BBPMCS auf dem OBDS unter Zulassung von Lücken für den Multi-Assay Benchmarkdatensatz (Knotenlabel „elements“ Element, „fconv“ Atomtypen, „fconv_groups“ Atomtypengruppen, „ph4_H“ und „ph4_noH“ Pharmakophor-Ebene unter bzw. ohne Berücksichtigung von Wasserstoffatomen, Zahl: Gap-Strafe).

A.9. Clusteringergebnisse von StruClus unter Verwendung von verschiedenen Repräsentationen und Ähnlichkeitsebenen für den SBDS



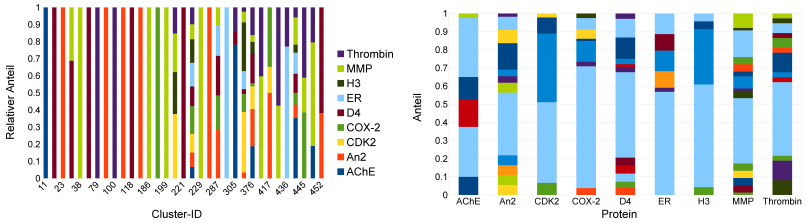
(a) Element-Label

(b) Element-Label



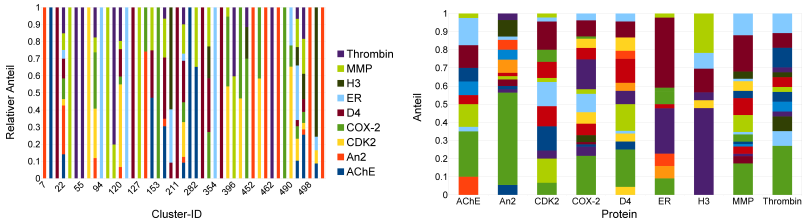
(c) fconv-Atomtypen

(d) fconv-Atomtypen



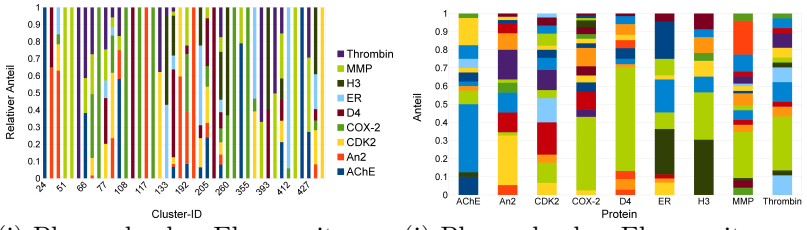
(e) fconv-Atomtypengruppen

(f) fconv-Atomtypengruppen



(g) Sybyl-Atomtypen

(h) Sybyl-Atomtypen



(i) Pharmakophor-Ebene mit Wasserstoffatomen

(j) Pharmakophor-Ebene mit Wasserstoffatomen

Abbildung A.17.: Ergebnisse von StruClus unter Verwendung von verschiedenen Repräsentationen und Ähnlichkeitsebenen für den SBDS. Die linke Abbildung stellt jeweils eine Cluster-zentrierte Analyse und die rechte Abbildung eine Target-Protein zentrierte Analyse dar. Bei der Target-Protein zentrierten Analyse stehen die Farben für die Cluster.

Target-Protein	Jarvis-Patrick	original StruClus	Elemente	fconv	fconv-Gruppen	Sybyl	ph4_noH
AChE	0,5209 0,4650 (10/22)	0,5439 0,4394 (6/20)	0,3726 0,2011 (10/16)	0,3077 0,1571 (11/15)	0,3914 0,2549 (6/15)	0,3620 0,1797 (10/17)	0,3572 0,1598 (12/23)
D4	0,7777 0,4552 (14/25)	0,5622 0,3809 (9/12)	0,2575 0,2277 (10/22)	0,3741 0,2970 (16/21)	0,5239 0,0448 (12/17)	0,5042 0,2391 (14/25)	0,4168 0,2200 (12/26)
An2	0,6380 0,5692 (10/35)	0,8385 0,4984 (14/18)	0,6010 0,2055 (9/17)	0,5550 0,2472 (11/19)	0,5403 0,0358 (12/14)	0,5569 0,2275 (13/21)	0,3754 0,1996 (13/25)
ER	0,7467 0,3657 (4/19)	0,5436 0,3444 (7/13)	0,4479 0,2467 (12/17)	0,3635 0,2167 (9/19)	0,3749 0,0259 (6/13)	0,4542 0,1850 (9/17)	0,3094 0,1987 (10/26)
MMP	0,6745 0,4882 (14/25)	0,5764 0,4242 (7/20)	0,4339 0,3036 (14/21)	0,3024 0,2839 (16/18)	0,5135 0,2591 (15/15)	0,4147 0,2785 (16/33)	0,4163 0,2360 (17/28)
H3	0,7596 0,1977 (10/14)	0,4073 0,2109 (7/7)	0,2198 0,1157 (7/11)	0,2344 0,1634 (7/14)	0,0765 0,2063 (5/14)	0,2141 0,2029 (6/13)	0,2105 0,1503 (8/17)
Thrombin	0,6581 0,2323 (13/18)	0,6882 0,3484 (11/12)	0,3450 0,1791 (9/16)	0,1510 0,2273 (10/19)	0,2868 0,0153 (13/16)	0,3562 0,1522 (13/15)	0,2520 0,1572 (14/19)
COX-2	0,7081 0,5355 (19/29)	0,7871 0,5376 (17/23)	0,3499 0,2868 (12/23)	0,5754 0,2514 (12/19)	0,3454 0,2943 (8/17)	0,4369 0,3338 (15/32)	0,5361 0,2252 (16/29)
CDK2	0,8222 0,3107 (31/21)	0,5884 0,3688 (9/10)	0,3635 0,1860 (11/16)	0,2585 0,1824 (13/17)	0,2109 0,0271 (5/14)	0,3243 0,1913 (12/17)	0,2765 0,1712 (14/22)

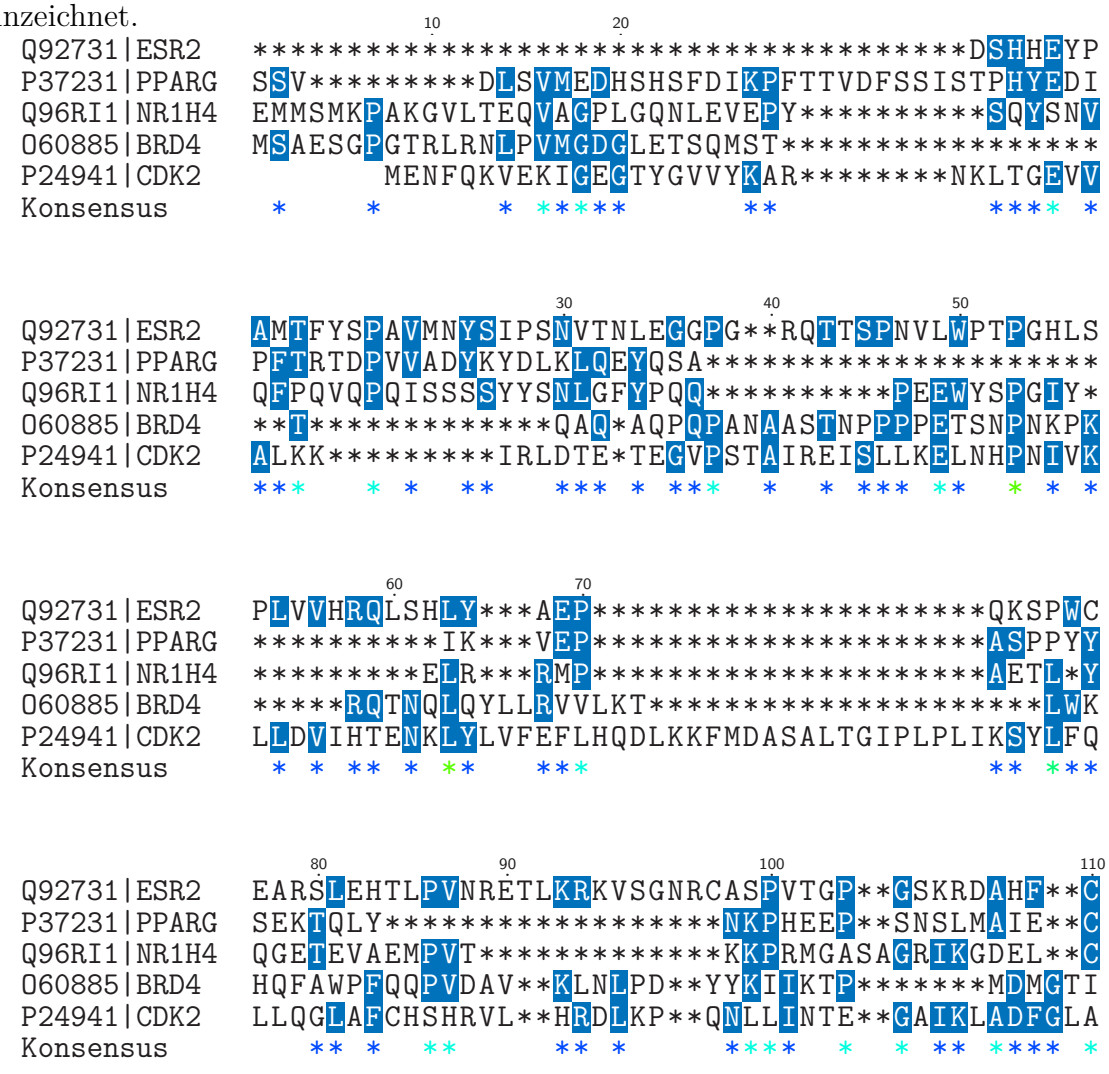
Tabelle A.4.: Ergebnisse von StruClus unter Verwendung von verschiedenen Repräsentationen und Ähnlichkeitsebenen für den SBDS. Angegeben ist die Gesamt-Reinheit (siehe Gleichung IV.2.8) für das Experiment (obere Reihe) und ein randomisiertes Clustering (mittlere Reihe). In Klammern ist die Anzahl der Cluster angegeben, zuerst für das Experiment und danach für ein randomisiertes Clustering. Proteinabkürzungen: Acetylcholinesterase (AChE), Angiotensin-II-Rezeptor (An2), Cyclin-abhängige Kinase 2 (CDK2), Cyclooxygenase-2 (COX-2), Dopamin D4-Rezeptor (D4), Estrogenrezeptor (ER), Histamin-H3-Rezeptor (H3), Matrix-Metalloproteinase 3 (MMP). Grün: höchster Wert, lila: niedrigster Wert. ph4_noH: Pharmakophor-Ebene ohne Berücksichtigung von Wasserstoffatomen.

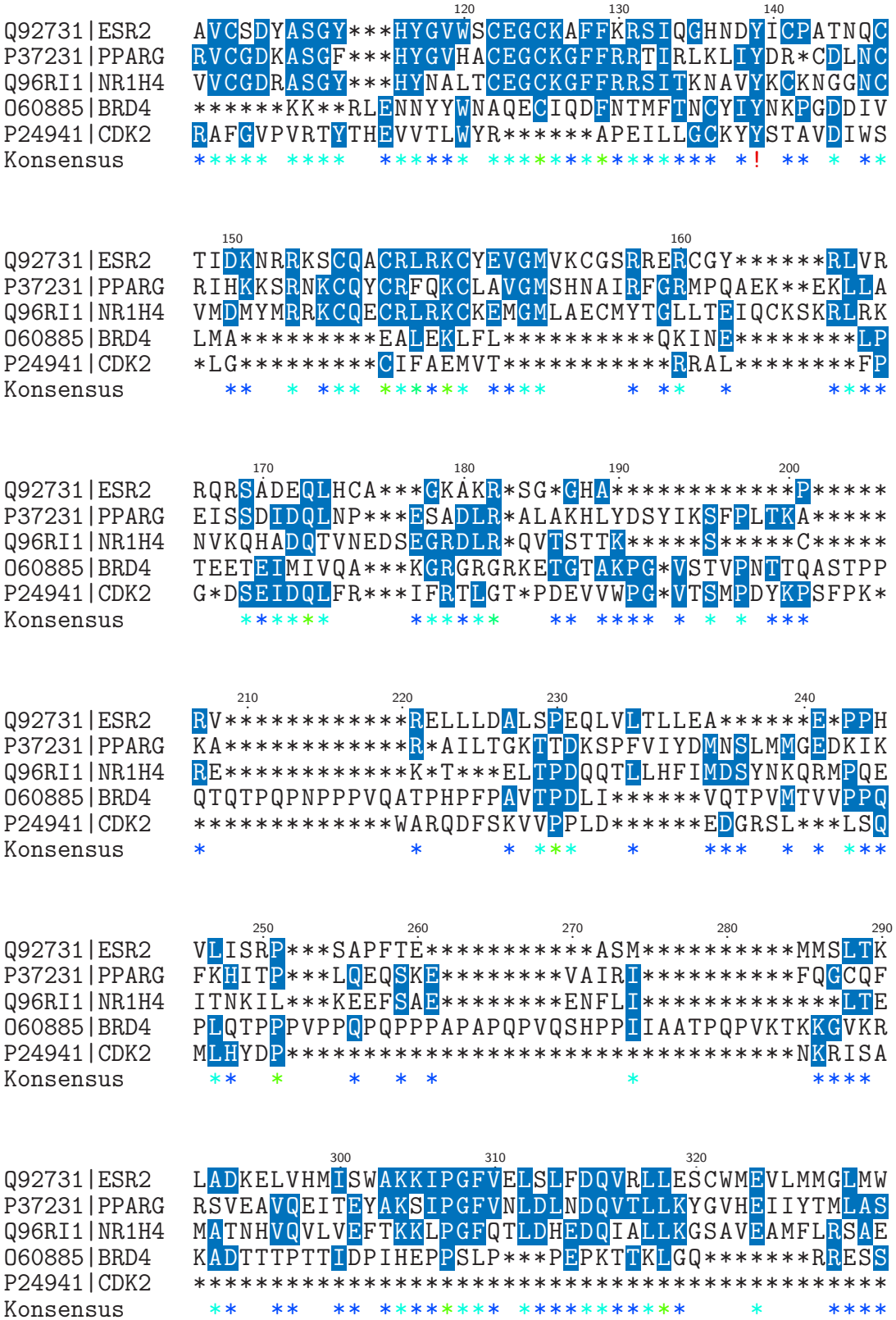
B. Die Scaffold- und Protein-Ebene der Ähnlichkeit

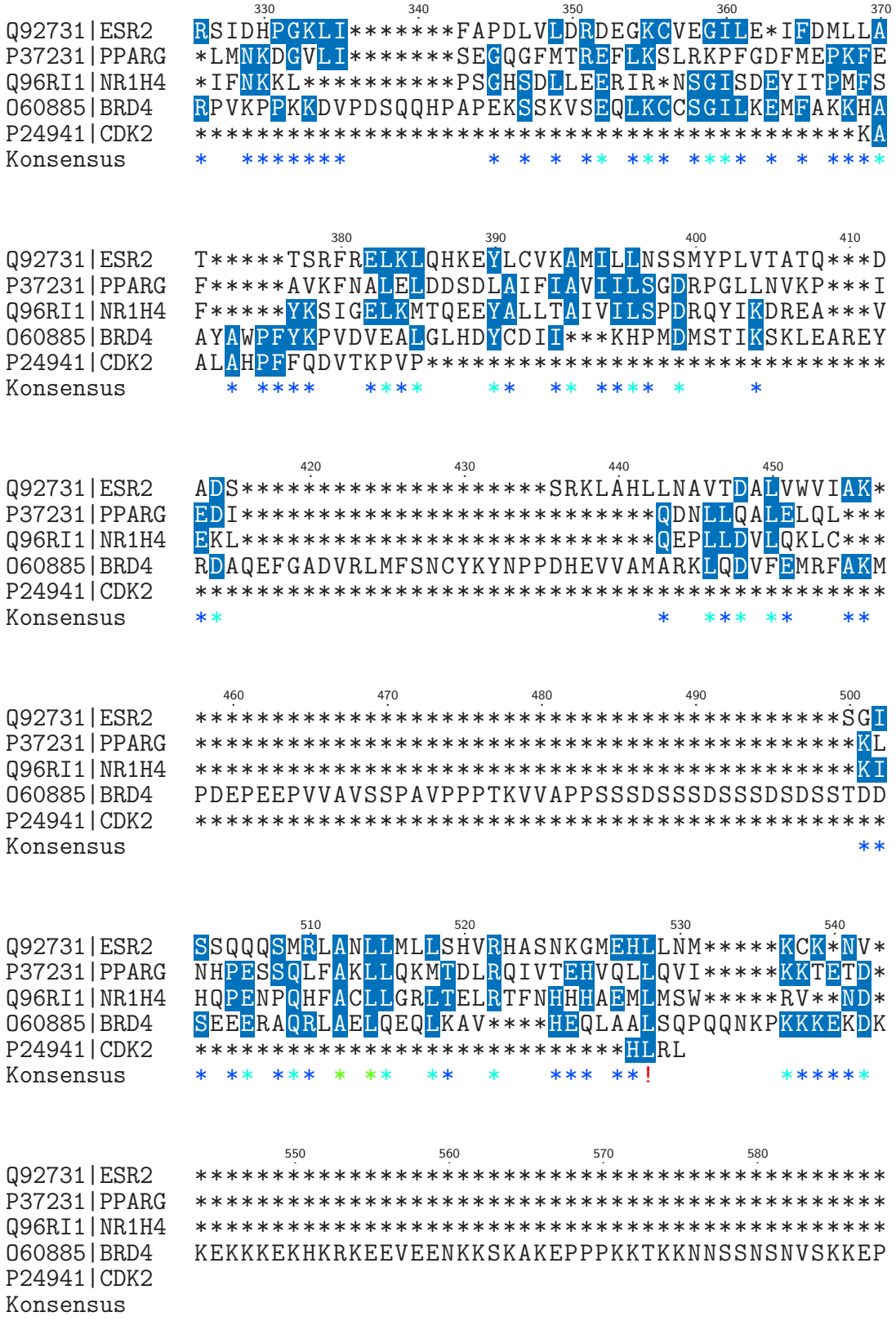
B.1. Alignment der Proteinstrukturen aus dem APSC

B.1.1. Globales Alignment mittels Clustal Omega

Ein globales Alignment der APSC-Proteine wurde erstellt mit Clustal Omega (Sievers *et al.* 2011; *Clustal Omega* 2018; *Clustal Omega - Multiple Sequence Alignment (Webserver)* 2018) und dargestellt mithilfe des Paketes *TEXshade* (Beitz 2000). Unter dem Konsensus-Eintrag wird das Ausmaß an Übereinstimmung in den Sequenzen mit einem Farbspektrum von blau (Übereinstimmung in zwei Sequenzen), über türkis (Übereinstimmung in drei Sequenzen) und grün (Übereinstimmung in vier Sequenzen) nach rot (Übereinstimmung in allen fünf Sequenzen) angegeben. Die Sequenznummerierung wurde von BRD4 übernommen. Eingefügte Lücken (*gaps*) innerhalb der Sequenz sind mit einem schwarzen Stern gekennzeichnet.




Q92731|ESR2
P37231|PPARG
Q96RI1|NR1H4
O60885|BRD4
P24941|CDK2
Konsensus


330 340 350 360 370
Q92731|ESR2 RSIDHPGKLI*******FAPDLVLDRDEGKCVEGILE*IFDMLLA
P37231|PPARG *LMNKDGVLI*******SEGQGFMTREFLKSLRKPFGDFMEPKFE
Q96RI1|NR1H4 *IFNKKL**********PSGHSDLLEERIR*NSGISDEYITPMFS
O60885|BRD4 RPVKPPKKDVPDSQQHPAPEKSSKVSEQLKCCSGILKEMFAKKHA
P24941|CDK2 **KA
Konsensus * ******* * * * ** *** **** * * ****
380 390 400 410
Q92731|ESR2 T*****TSRFRELKLQHKEYLCVKAMILLNSSMYPLVTATQ***D
P37231|PPARG F*****AVKFNALELDDSDLAIFIAVIILSGDRPGLLNVKP***I
Q96RI1|NR1H4 F*****YKSIGELKMTQEEYALLTAIVILSPDRQYIKDREA***V
O60885|BRD4 AYAWPFYKPVDVEALGLHDYCDII***KHPMDMSTIKSKLEAREY
P24941|CDK2 ALAHPFFQDVTKPVP****************************
Konsensus * **** **** ** ** **** * *
420 430 440 450
Q92731|ESR2 ADS*******************SRKLAHLLNAVTDALVWVIAK*
P37231|PPARG EDI*****************************QDNLLQALELQL***
Q96RI1|NR1H4 EKL*****************************QEPLLDVLQKLC***
O60885|BRD4 RDAQEFGADVRLMFSNCYKYNPPDHEVVAMARKLQDVFEMRFAKM
P24941|CDK2 **
Konsensus ** * *** ** **
460 470 480 490 500
Q92731|ESR2 **SGI
P37231|PPARG **KL
Q96RI1|NR1H4 **KI
O60885|BRD4 PDEPEEPVVAVSSPAVPPPTKVVAPPSSSDSSSDSSSDSDSSTDD
P24941|CDK2 **
Konsensus **
510 520 530 540
Q92731|ESR2 SSQQQSMRLANLLMLLSHVRHASNKGMEHLLNM*****KCK*NV*
P37231|PPARG NHPESSQLFAKLLQKMTDLRQIVTEHVQLLQVI*****KKTETD*
Q96RI1|NR1H4 HQPENPQHFACLLGRLTELRTFNHHHAEMLMSW*****RV**ND*
O60885|BRD4 SEEERAQRLAELQEQLKAV****HEQLAALSQPQQNKPKKKEKDK
P24941|CDK2 ***************************HLRL
Konsensus * ** *** * ** ** * *** **! ******
550 560 570 580
Q92731|ESR2 **
P37231|PPARG **
Q96RI1|NR1H4 **
O60885|BRD4 KEKKKEKHKRKEEVEENKKSKAKEPPPKKTKKNNSSNSNVSKKEP
P24941|CDK2
Konsensus

```
                 590         600         610         620         630
Q92731|ESR2     ********************************************
P37231|PPARG    ********************************************
Q96RI1|NR1H4    ********************************************
O60885|BRD4     APMKSKPPPTYESEEEDKCKPMSYEEKRQLSLDINKLPGEKLGRV
P24941|CDK2
Konsensus


                        640         650         660         670
Q92731|ESR2     *******************VPVYD*******************
P37231|PPARG    *******************MSLHP*******************
Q96RI1|NR1H4    *******************HKFTP*******************
O60885|BRD4     VHIIQSREPSLKNSNPDEIEIDFETLKPSTLRELERYVTSCLRKK
P24941|CDK2
Konsensus                              * *


                 680         690         700         710         720
Q92731|ESR2     ******LLLEMLNAHVLRGCKSSITGSEC*SPAEDSKSKEGSQNP
P37231|PPARG    ******LLQEIYKDLY
Q96RI1|NR1H4    ******LLCEIWDVQ
O60885|BRD4     RKPQAEKVDVIAGSSKMKGFSSSESESSSESSSSDSEDSETEMAP
P24941|CDK2
Konsensus             ** **       *  **   *   *   **   *    *


                        730
Q92731|ESR2     QSQ
P37231|PPARG
Q96RI1|NR1H4
O60885|BRD4     KSKKKGH
P24941|CDK2
Konsensus        *


                X  nicht konserviert
                X  ≥ 50% konserviert
```

B.1.2. Phylogenetischer Baum

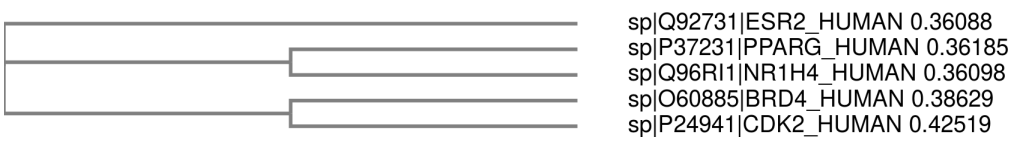


Abbildung B.1.: Phylogenetischer Baum der Proteine aus dem APSC. Erstellt nach dem Clustal Omega Alignment von *Clustal Omega - Multiple Sequence Alignment (Webserver)* (2018).

B.1.3. Ergebnisse eines lokalen Alignments mittels BLAST

Es wurde ebenfalls ein lokales Alignment der APSC-Proteine mithilfe von BLAST (*BLAST - Basic Local Alignment Search Tool* 2018; Altschul *et al.* 1990) durchgeführt. Die Ergebnisse sind in den folgenden Tabellen aufgeführt.

gefunden: \ gesucht:	BRD4	PPARγ	CDK2	ESR2	FXR
BRD4	2686/2686	18,9/37,3		21,9/21,9	18,1/35,4
PPARγ	18,9/37,3	1049/1049	17,7/17,7	84,7/84,7	148/148
CDK2		17,7/17,7	616/616		
ESR2	21,6/41,6	84,3/84,3		1101/1101	107/107
FXR		148/148		107/107	1021/1021

Tabelle B.1.: BLAST Ergebnis für APSC-Proteine: maximaler (erster Wert) und totaler Score (zweiter Wert).

gefunden: \ gesucht:	BRD4	PPARγ	CDK2	ESR2	FXR
BRD4	100	4		19	4
PPARγ	2	100	8	36	71
CDK2		5	100		
ESR2	10	42		100	70
FXR		71		64	100

Tabelle B.2.: BLAST Ergebnis für APSC-Proteine: Suchsequenz-Abdeckung in Prozent.

gefunden: \ gesucht:	BRD4	PPARγ	CDK2	ESR2	FXR
BRD4	0,0	3,1		0,33	4,4
PPARγ	8,1	0,0	3,7	$6*10^{-21}$	10^{-42}
CDK2		6,3	0,0		
ESR2	1,2	$7*10^{-21}$		0,0	$2*10^{-28}$
FXR		10^{-42}		10^{-28}	0,0

Tabelle B.3.: BLAST Ergebnis für APSC-Proteine: Erwartungswert (E-Wert).

gefunden: \ gesucht:	BRD4	PPARγ	CDK2	ESR2	FXR
BRD4	100	32		22	33
PPARγ	38	100	27	30	31
CDK2		27	100		
ESR2	23	30		100	27
FXR		31		27	100

Tabelle B.4.: BLAST Ergebnis für APSC-Proteine: Identität in Prozent.

B.2. Analyse des Chromon-Scaffolds als erweitertes privilegiertes Scaffold

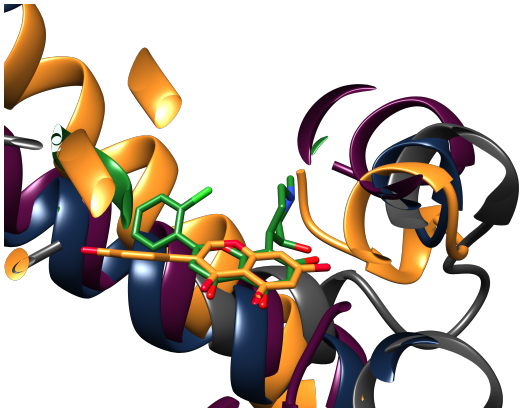
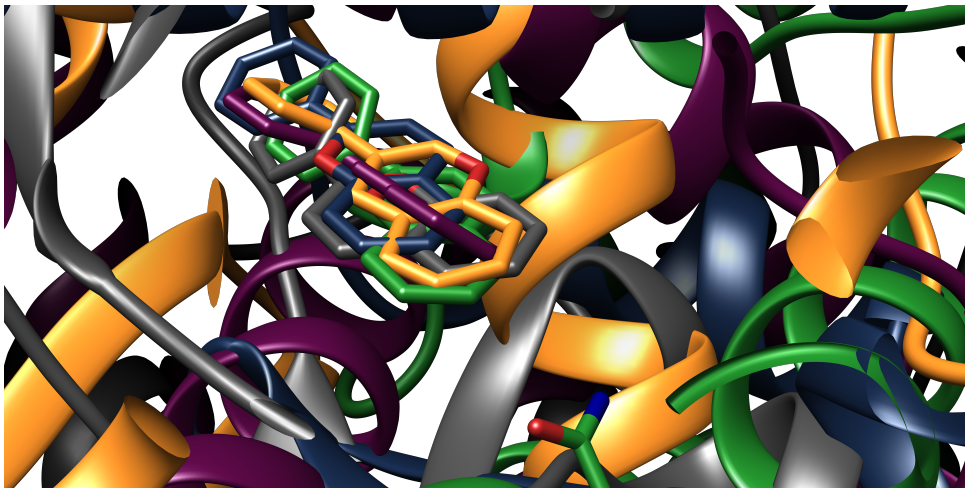
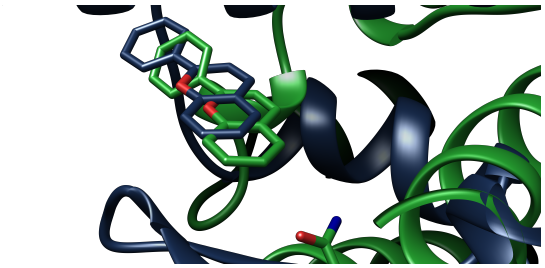


Abbildung B.2.: Strukturen von BRD4 (grün, PDB-ID: 4O71), PPARγ (blau, PDB-ID: 1NYX), CDK2 (grau, PDB-ID: 1GZ8), ESR2 (orange, PDB-ID: 1QKM) sowie NR1H4 (weinrot, PDB-ID: 3BEJ) mit dem Fokus auf den Liganden der BRD4- und ESR2-Komplexstrukturen.

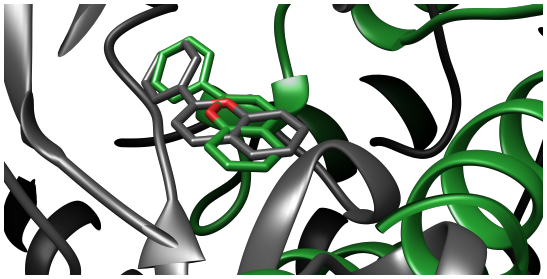
Die Scaffolds der kokristallisierten Liganden der ESR2- und BRD4-Komplexstrukturen sind gut aligniert, wenn die Kristallstrukturen manuell im Hinblick auf ähnliche Sekundärstrukturen überlagert werden.



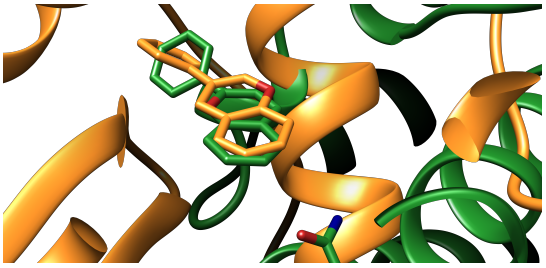
(a) Alignment aller APSC Proteine



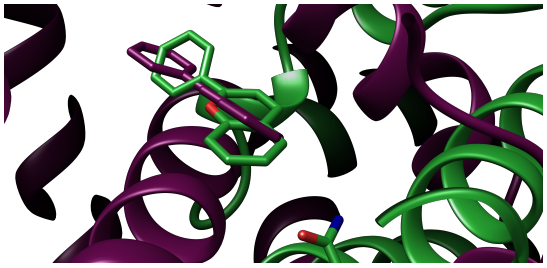
(b) BRD4 (grün) und PPARγ (blau)



(c) BRD4 (grün) und CDK2 (grau)



(d) BRD4 (grün) und ESR2 (orange)



(e) BRD4 (grün) und NR1H4 (weinrot)

Abbildung B.3.: Alignment aller APSC-Proteine (a) und von BRD4 (grün, PDB-ID: 4O71) mit PPARγ (blau, PDB-ID: 3SZ1) (b), mit CDK2 (grau, PDB-ID: 2DUV) (c), mit ESR2 (orange, PDB-ID: 1QKM) (d) sowie mit NR1H4 (weinrot, PDB-ID: 3BEJ) (e) mittels IsoMIF basierend auf dem 2-Phenyl-4H-chromen bzw. analogen Scaffold.

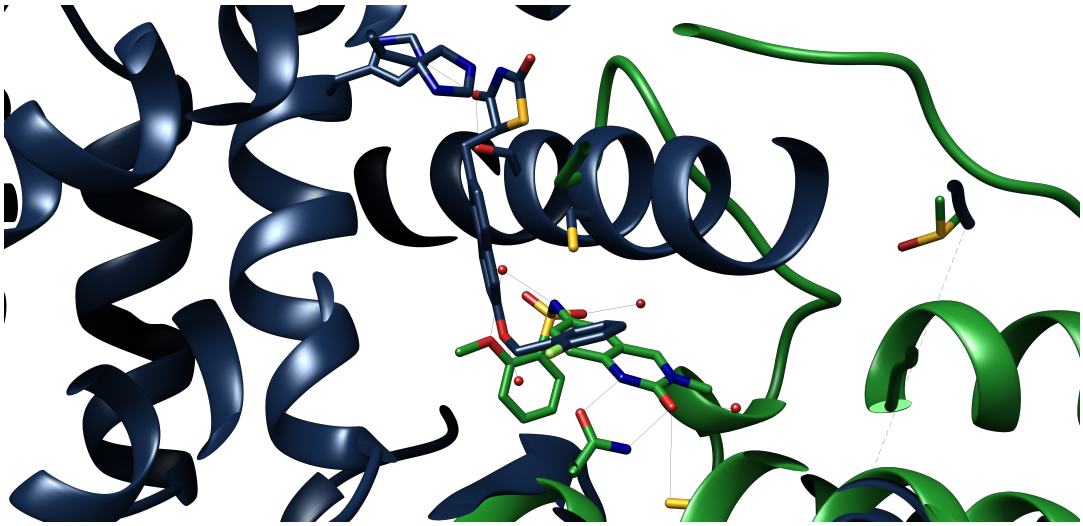
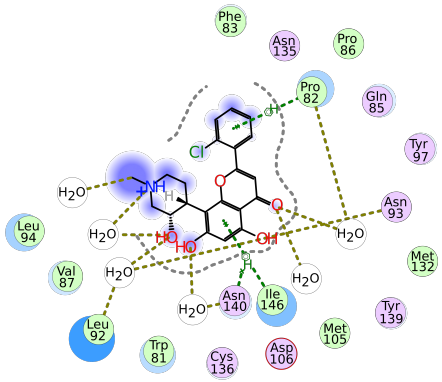
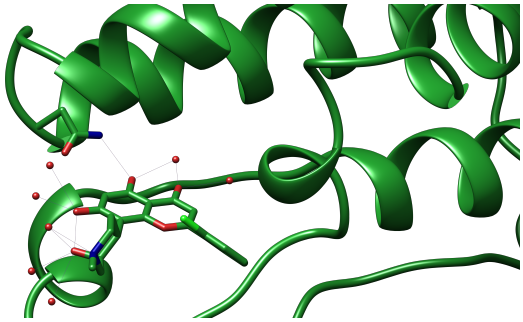


Abbildung B.4.: IsoMIF-basiertes Alignment der Komplexstrukturen von BRD4 (grün, PDB-ID: 4E96) mit dem selektivem Liganden PFI-1 und PPARγ (blau, PDB-ID: 3B0Q) mit dem selektivem Liganden Netoglitazon. Wasserstoffbrücken sind als lila Striche gekennzeichnet.

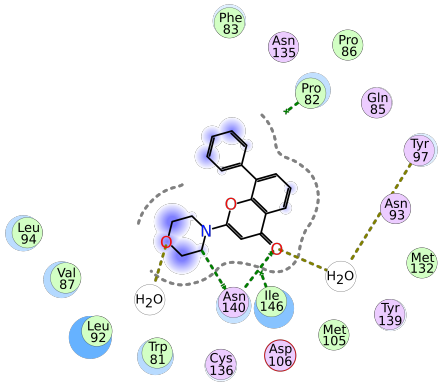
In Abbildung B.4 ist zu erkennen, dass die beiden selektiven Liganden Netoglitazon (PPARγ) und PFI-1 (BRD4) keine Protein übergreifenden Wasserstoffbrückenbindungen ausbilden. Es bestehen nur Wasserstoffbrückenbindungen zwischen dem Liganden und der dazugehörigen Proteinstruktur.



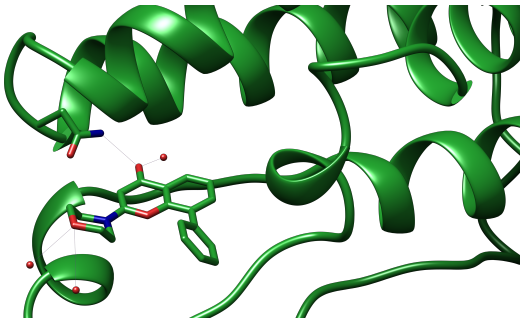
(a) Interaktion von BRD4(1) mit Chromon-Scaffold (PDB-ID: 4O71)



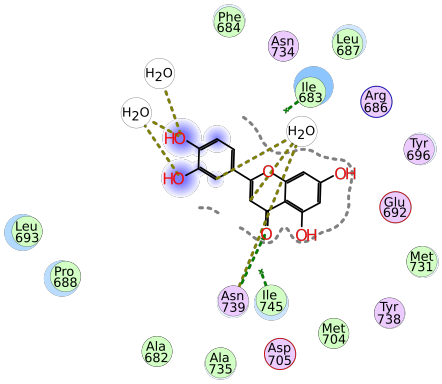
(b) BRD4(1) mit Chromon-Scaffold (PDB-ID: 4O71)



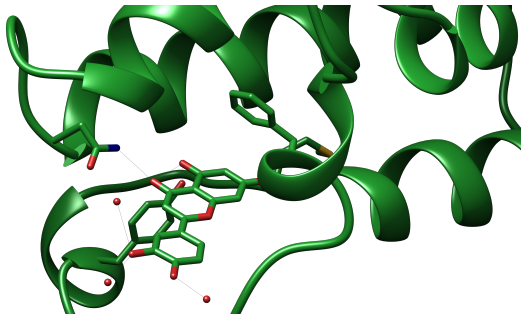
(c) Interaktion von BRD4(1) mit Chromon-Scaffold (PDB-ID: 4CFK)



(d) BRD4(1) mit Chromon-Scaffold (PDB-ID: 4CFK)

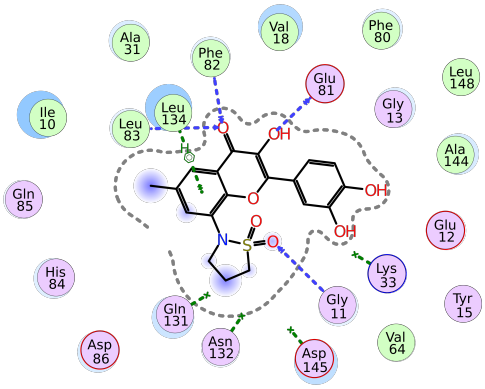


(e) Interaktion von PBRM1(5) mit Chromon-Scaffold (PDB-ID: 5II2)

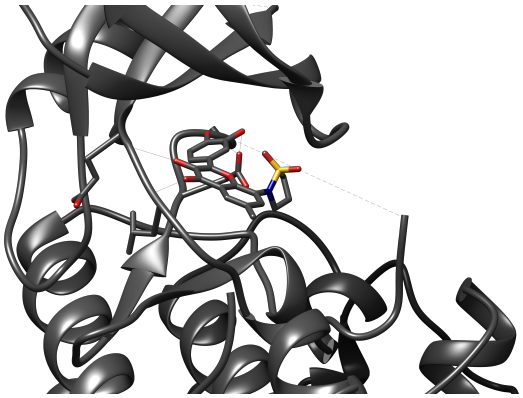


(f) PBRM1(5) mit Chromon-Scaffold (PDB-ID: 5II2)

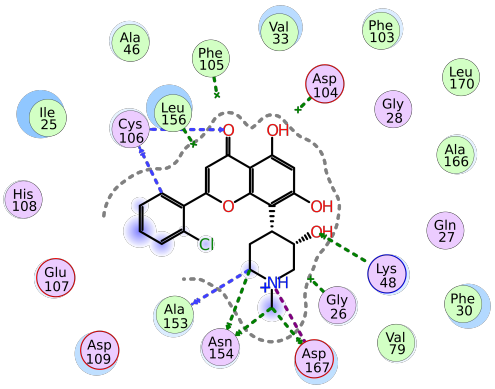
Abbildung B.5.: Interaktionen des Chromon-Scaffolds mit BRDs. Die linke Abbildung zeigt das Interaktionsmuster in 2D und wurde mithilfe von MOE (*Molecular Operating Environment (MOE), 2013.08* 2013) erstellt. Die rechte Abbildung wurde mit Chimera erstellt und zeigt die 3D-Orientierung des Scaffolds, Wasserstoffbrückenbindungen sind als lila Striche gekennzeichnet. PBRM1: Protein Polybromo-1.



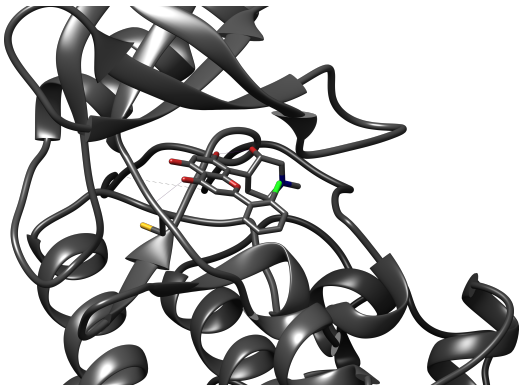
(a) Interaktion von CDK2 mit Chromon-Scaffold (PDB-ID: 2DUV)



(b) CDK2 mit Chromon-Scaffold (PDB-ID: 2DUV)



(c) Interaktion von CDK9 mit Chromon-Scaffold (PDB-ID: 3BLR)



(d) CDK9 mit Chromon-Scaffold (PDB-ID: 3BLR)

Abbildung B.6.: Interaktionen des Chromon-Scaffolds mit CDK2 (oben) und CDK9 (unten). Die linke Abbildung zeigt das Interaktionsmuster in 2D und wurde mithilfe von MOE (*Molecular Operating Environment (MOE), 2013.08* 2013) erstellt. Die rechte Abbildung wurde mit Chimera erstellt und zeigt die 3D-Orientierung des Scaffolds, Wasserstoffbrückenbindungen sind als lila Striche gekennzeichnet.

B.3. Ergebnisse des initialen Screens

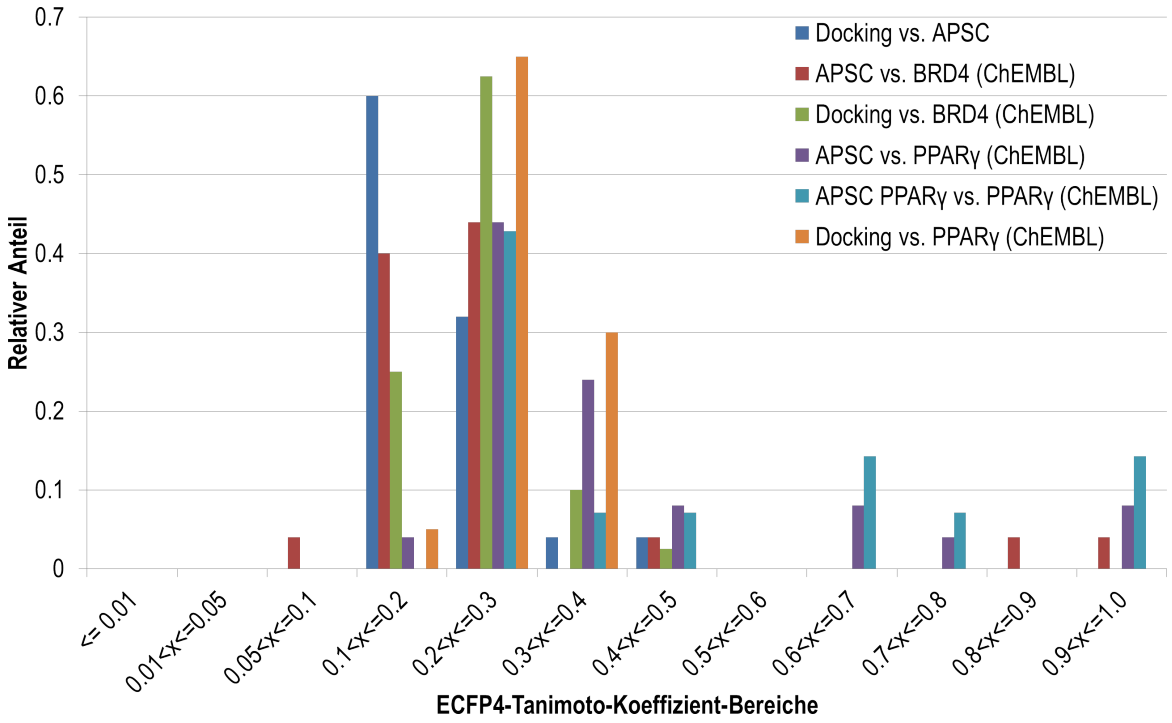
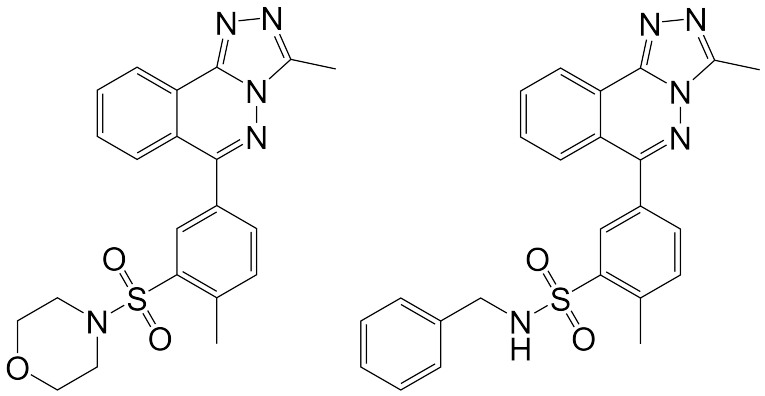


Abbildung B.7.: Ähnlichkeitsanalyse basierend auf dem ECFP4-TC von BRD4-Hits aus dem APSC und dem Docking-Ansatz sowie Vergleich zu bekannten BRD4- und PPARγ-Liganden. Der ECFP4-TC wurde mithilfe von Pipeline Pilot berechnet und der maximale Wert je Verbindung im Histogramm aufgetragen. Der experimentellen Analyse wurden drei Positivkontrollen (JQ1, Verbindungen **24** und **25** (Abbildung B.8)) hinzugefügt. Diese weisen einen ECFP4-TC von mindestens 0,8 zu mindestens einem bekannten BRD4-Liganden auf. Wobei JQ1 nicht im gezeigten Histogramm enthalten ist.

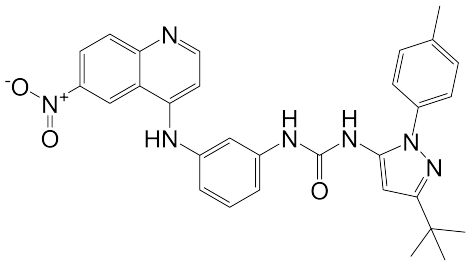
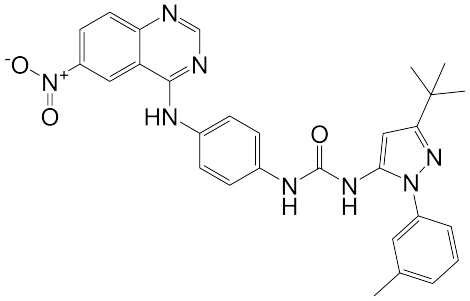
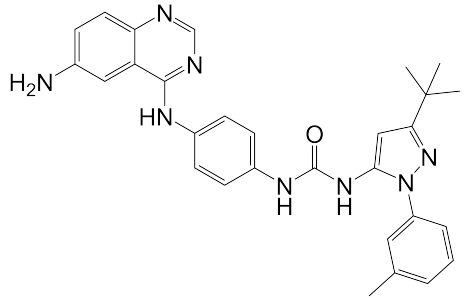
Die Ähnlichkeit der mittels APSC gefundenen Hits zu den Hits basierend auf dem Docking-Ansatz sollte ermittelt werden. Zudem sollte die Ähnlichkeit der jeweiligen Hits zu bekannten BRD4- und PPARγ-Liganden analysiert werden. Dazu wurde der ECFP4 der Strukturen erstellt und der TC mittels Pipeline Pilot (*BIOVIA Pipeline Pilot: BIOVIA Workbook* 2015) darüber berechnet. Ein Histogramm der maximalen ECFP4-TC-Werte befindet sich in Abbildung B.7. Die Hits des Docking-Ansatzes zeigen zu den Hits des APSC eine geringe Ähnlichkeit. Die Ähnlichkeit der APSC-Hits zu bekannten BRD4-Inhibitoren ist ebenfalls gering mit zwei Ausnahmen, welche sich auf die beiden Positivkontrollen **24** und **25** (Strukturen siehe Abbildung B.8) zurückführen lassen. Die Docking-basierten Hits sind sowohl zu bekannten BRD4- als auch zu bekannten PPARγ-Liganden unähnlich. Unter den APSC-Hits sind einige Verbindungen ähnlich zu PPARγ-Liganden, insbesondere, wenn das Subset an APSC-Hits betrachtet wird, welches auf zu PPARγ-Liganden änlichen Liganden beruht (APSC PPARγ).

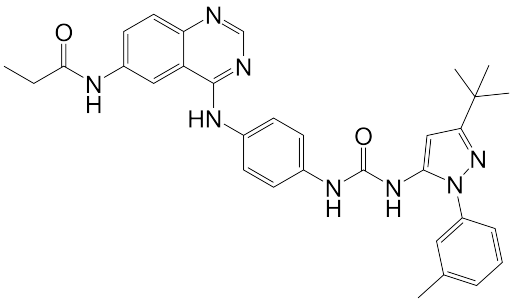
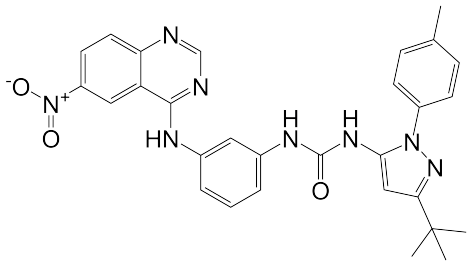
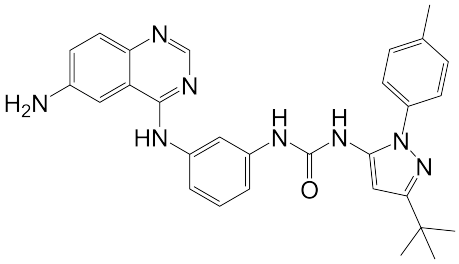
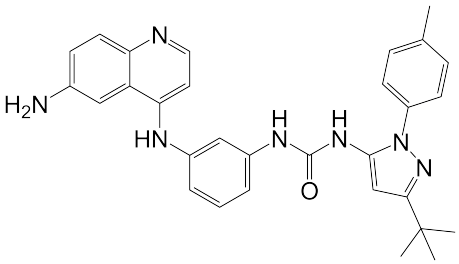
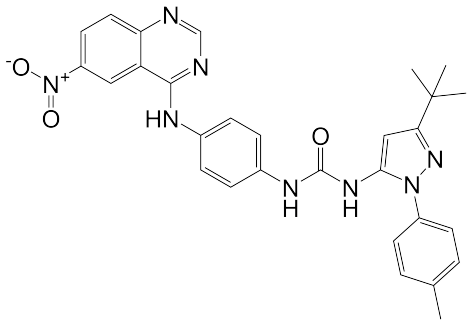


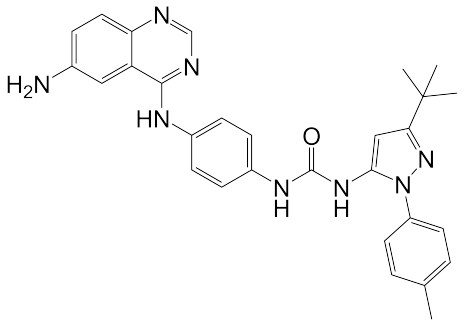
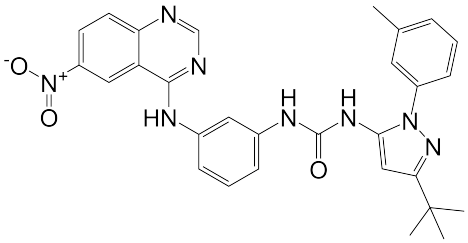
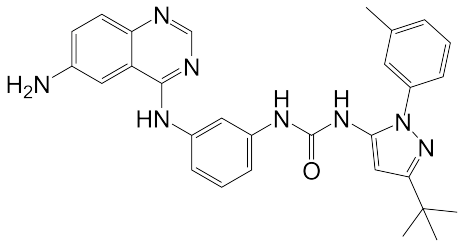
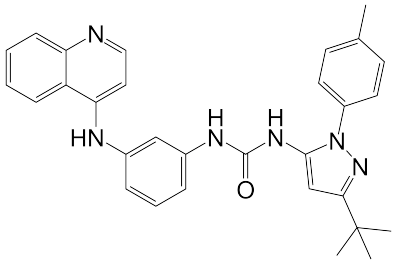
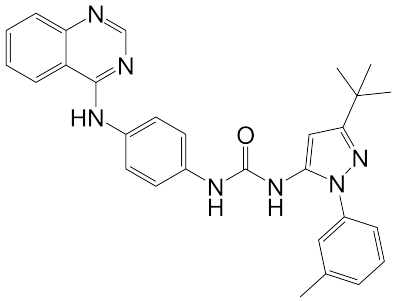
(a) **24** (b) **25**

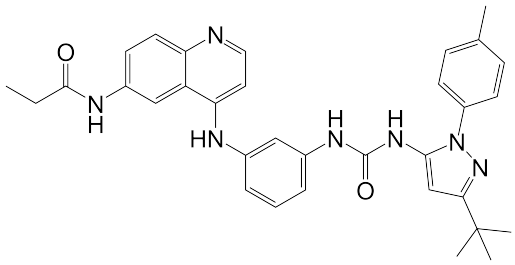
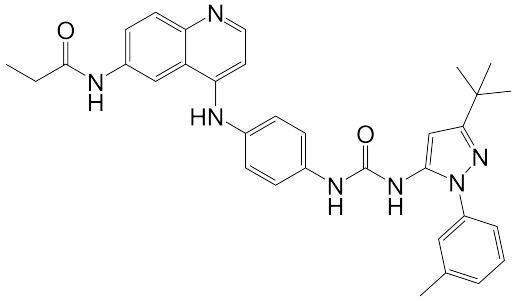
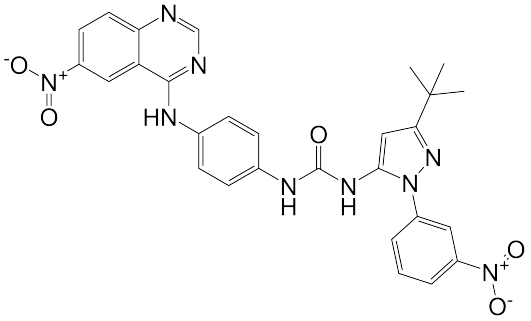
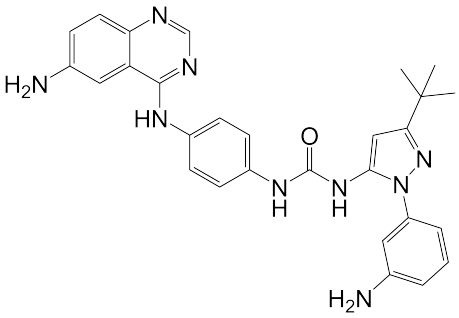
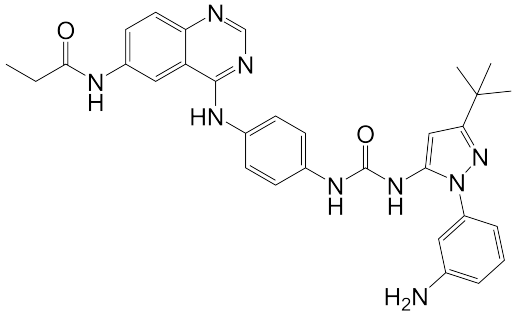
Abbildung B.8.: Strukturen der Positivkontrollen **24** (a) und **25** (a).

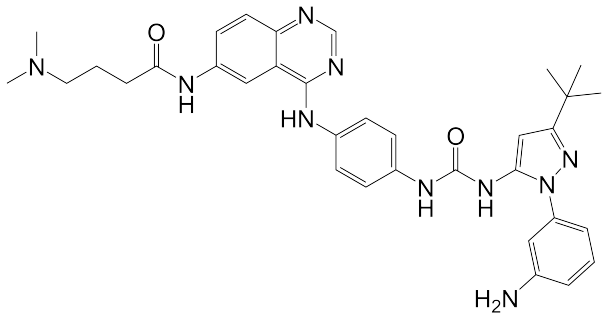
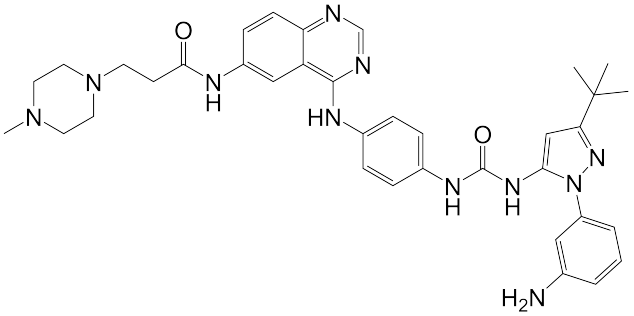
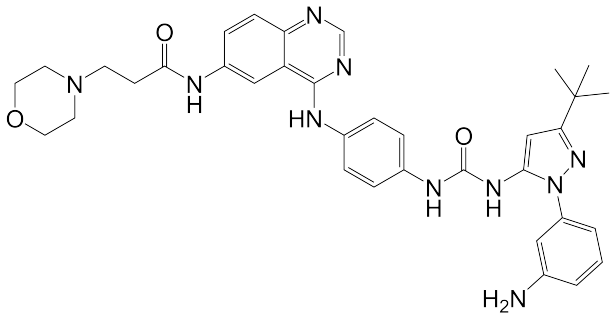
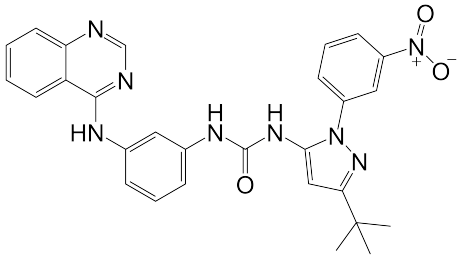
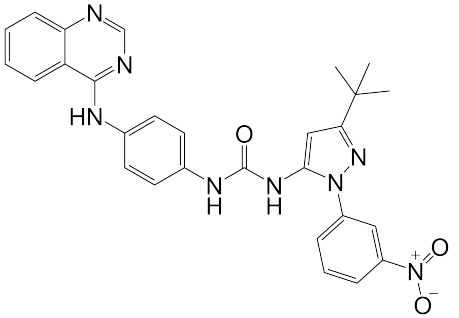
B.4. Biochemische Charakterisierung einer Substanzserie zur Inhibition von EGFR-Varianten

Substanz-ID	Struktur	Ø $IC_{50} \pm \sigma$ in μM
RL024		n. a.
RL044		n. a.
RL045		n. a.

Substanz-ID	Struktur	Ø IC_{50} ± σ in μM
RL046		n. a.
RL047		n. a.
RL048		n. a.
RL051		n. a.
RL059		n. a.

Substanz-ID	Struktur	Ø IC_{50} ± σ in μM
RL060		n. a.
RL061		n. a.
RL062		n. a.
RL063		n. a.
RL070		n. a.

Substanz-ID	Struktur	Ø IC_{50} ± σ in μM
RL077		n. a.
RL078		5,36 ± 0,18
RL100		n. a.
RL101		n. a.
RL102		n. a.

Substanz-ID	Struktur	Ø IC_{50} ± σ in μM
RL103		n. a.
RL104		n. a.
RL105		n. a.
RL127		> 10
RL129		7,92 ± 0,07

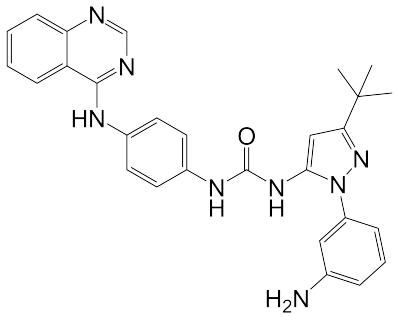
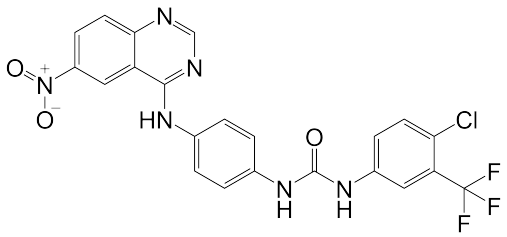
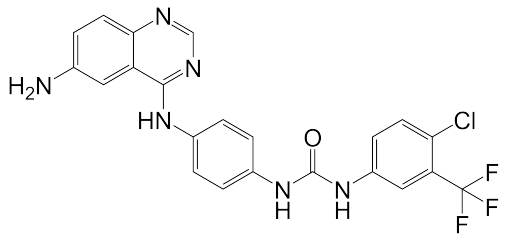
Substanz-ID	Struktur	Ø IC_{50} ± σ in μM
RL130		n. a.
RL132		n. a.
RL178		n. a.

Tabelle B.5.: Ergebnisse der EGFR-Inhibitor-Serie für BRD4. n. a.: nicht auswertbar bei einer Konzentrationsreihe bis maximal 10 μM.

B.5. Analyse der zellulären Effekte der BRD4-Inhibitoren aus der EGFR-Inhibitor-Serie

Mutation	$\text{Rang}_{sensitiv}$	$\text{Rang}_{insensitiv}$	Mutation	$\text{Rang}_{sensitiv}$	$\text{Rang}_{insensitiv}$
MT-CYB	1	7	*WRNIP1*	4	28
TP53	2	1	*DIP2C*	4	27
HERC2	3	16	*PAPPA2*	4	27
FMN2	3	15	*MYH1*	4	24
MT-ND1	3	15	*PRUNE2*	4	23
MT-ND5	3	4	*DYNC2H1*	4	20
MUC16	3	3	*MT-CO2*	4	16
TTN	3	2	*MT-CO3*	4	16
FAM84A	**4**	**35**	*VCAN*	4	15
BCAS3	**4**	**31**	*BIRC6*	4	14
AHNAK	4	20	*MT-ND4*	4	14
KIAA1462	**4**	**30**	*SYNE1*	4	14
AKR1C4	4	29	*FLG*	4	12
KIAA1244	4	28	*FAT3*	4	8
SORCS3	4	28	*LRP1B*	4	6
TLN2	4	28			

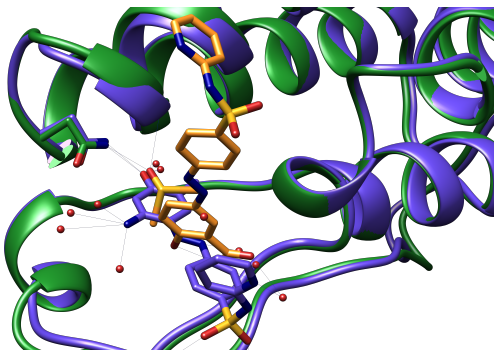
Tabelle B.6.: Gene die in mindestens der Hälfte der Zelllinien mit besonders sensitiver Reaktion auf duale EGFR-BRD4-Inhibition mutiert sind sowie deren Rang nach Häufigkeit sortiert für die sensitiven und weniger sensitiven (insensitiven) Zelllinien. Die Mutationsdaten entstammen aus *Cancer Cell Line Encyclopedia* (2018) (Cancer Cell Line Encyclopedia Consortium und Genomics of Drug Sensitivity in Cancer Consortium 2015). Fett markiert sind die drei Gene, deren Rang sich am meisten zwischen besonders sensitiven und insensitiven Zelllinien unterscheidet.

B.6 Kokristallstrukturen von BRD4

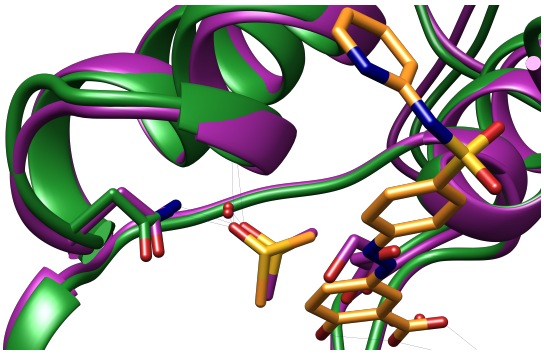
B.6.1. Datensammlungs- und Verfeinerungsstatistiken der BRD4(1)-Komplexstrukturen

Datensammlung	BRD4:**2** (6G0D)	BRD4:**3** (6G0E)	BRD4:**4** (6G0F)	BRD4:**22** (6G0G)	BRD4:**21** (6G0H)
Wellenlänge (Å)	0,999850	0,997100	0,919740	0,9999	1,7699
Röntgenquelle	SLS X10SA	SLS X10SA	SLS X10SA	SLS X10SA	SLS X10SA
Raumgruppe	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12$	$P2_12_12$
Auflösung (Å)	39,79-1,31 (1,36-1,31)	40,14-1,61 (1,67-1,61)	39,26-1,62 (1,68-1,62)	37,0-1,48 (1,53-1,48)	37,04-1,91 (1,98-1,91)
Einheitszelle					
a, b, c (Å)	38,29; 42,87; 79,58	34,48; 46,75; 78,34	37,53; 44,33; 78,52	39,32; 108,18; 30,23	39,43; 108,10; 30,22
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
R_{merge} (%)	7,6 (143,5)	8,2 (132,4)	6,9 (50,3)	7,6 (89,3)	8,5 (54,4)
I/σ	20,05 (1,70)	17,71 (1,93)	19,45 (2,92)	19,61 (2,24)	12,24 (0,82)
Vollständigkeit (%)	98,9 (96,9)	99,9 (98,3)	99,9 (99,7)	100 (99,9)	99,1 (96,4)
Verfeinerung					
R_{work}/R_{free}	0,2023/0,23564	0,2332/0,2608	0,1845/0,2294	0,1928/0,2195	0,1844/0,2260
Anzahl an Atomen					
Protein	1086	1094	1079	1144	1093
Ligand	27	47	29	39	37
Wasser	145	47	100	126	87
B-Faktor	15,10	35,16	23,4	18,4	24,5
RMS Abweichung					
Bindungslängen (Å)	0,006	0,003	0,008	0,008	0,006
Bindungswinkel (°)	0,793	0,555	1,117	0,98	0,84
Ramachandran unerlaubte Regionen (%)	0,0	0,0	0,0	0,0	0,0

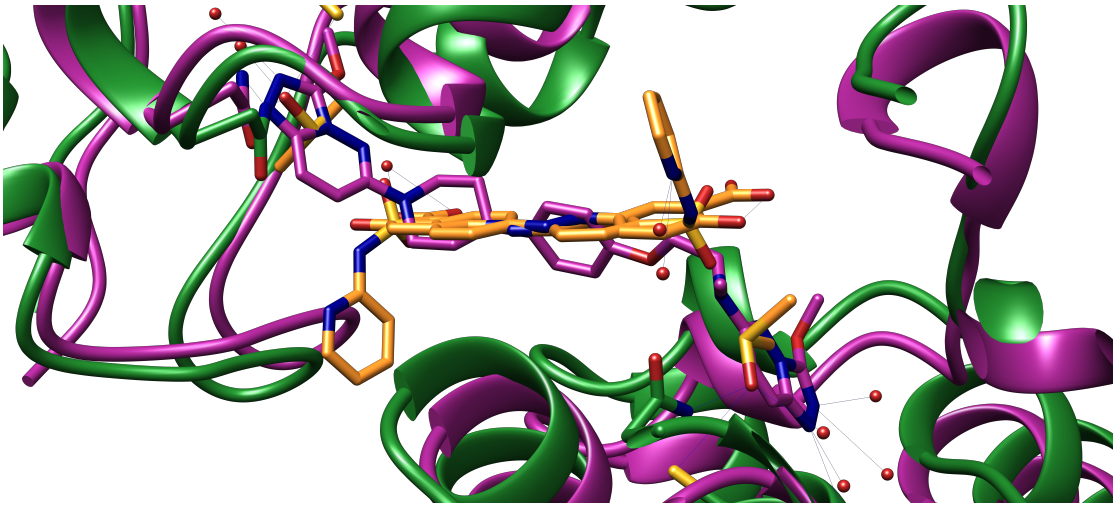
Tabelle B.7.: Datensammlungs- und Verfeinerungsstatistiken der BRD4-Komplexstrukturen mit den Verbindungen **2**, **3**, **4**, Sulfasalazin (**22**) und SSLH01 (**21**). Die Werte in Klammern beziehen sich jeweils auf die höchste Auflösungsschale.



(a) Sulfsalazin und MS436



(b) Sulfsalazin und DMSO



(c) Sulfsalazin und biBET

Abbildung B.9.: Alignment der BRD4-Sulfasalazin-Komplexstruktur (grün) mit relevanten weiteren Komplexstrukturen: BRD4-MS436-Komplexstruktur (PDB-ID: 4NUD) (a), BRD4-DMSO-Komplexstruktur (PDB-ID: 4IOR) (b) und BRD4-biBET-Komplexstruktur (PDB-ID: 5AD3) (c). In (c) wurde die BRD4-Sulfasalazin-Struktur als Dimer basierend auf Kristallkontakten verwendet. Sulfasalzin ist orange eingefärbt.

B.7. BRD4(1)-Sequenzen und Eigenschaften

B.7.1. BRD4(1)-Sequenz mit His-Tag

MHHHHHHSSGVDLGTENLYFQSMNPPPPETSNPNKPKRQTNQLQYLLRVVLKTLWKHQFA
WPFQQPVDAVKLNLPDYYKIIKTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYN
KPGDDIVLMAEALEKLFLQKINELPTEE

Molekülmasse: 17,5 kDa

Berechneter, molarer Extinktionskoeffizient* bei 280 nm (*ExPASy ProtParam Tool* 2018; Gasteiger *et al.* 2005): 28420 $M^{-1} \cdot cm^{-1}$

* Cystein-Reste als reduziert angenommen

B.7.2. BRD4(1)-Sequenz ohne His-Tag

`SMNPPPPETSNPNKPKRQTNQLQYLLRVVLKTLWKHQFAWPFQQPVDAVKLNLPDYYKII`
`KTPMDMGTIKKRLENNYYWNAQECIQDFNTMFTNCYIYNKPGDDIVLMAEALEKLFLQKI`
`NELPTEE`
Molekülmasse: 15,08 kDa
Berechneter, molarer Extinktionskoeffizient* bei 280 nm (*ExPASy ProtParam Tool* 2018; Gasteiger *et al.* 2005): 26930 $M^{-1} \cdot cm^{-1}$
* Cystein-Reste als reduziert angenommen

B.8. Zusammensetzungen von Puffern

B.8.1. Lysis-Puffer

HEPES pH 7,5	50 mM
Natriumchlorid	500 mM
Glycerol	5 %
TCEP	0,5 mM
Imidazol	5 mM

B.8.2. Größenausschluss-Puffer

HEPES pH 7,5	10 mM
Natriumchlorid	150 mM
TCEP	0,5 mM

B.8.3. TSA-Puffer

HEPES pH 7,5	10 mM
Natriumchlorid	150 mM

B.8.4. Reservoirlösungen

Inhibitor 2	Natriumfluorid	0,2 M
	PEG 3350	20 %
Inhibitoren 3, 21 und 22	Natriumacetat wasserfrei pH 4,6	0,1 M
	Natriumformiat	3,5 M
Inhibitor 4	Natriumcitrat	1,6 M

Danksagungen

Ich bedanke mich bei Prof. Dr. Daniel Rauh für die Übernahme des Erstgutachtens und die Möglichkeit die Experimente in seinen Laboren durchzuführen. Prof. Dr. Daniel Summerer danke ich für die Übernahme des Zweitgutachtens.

Prof. Dr. Petra Mutzel gilt mein Dank für die sehr angenehme und erfolgreiche Zusammenarbeit. Für eine gute Zusammenarbeit und tatkräftige Unterstützung bedanke ich mich ebenfalls bei Dr. Sonja Sievers vom COMAS und ihrem Team sowie Dr. Ingrid Vetter vom MPI Dortmund. Dr. Andreas Brunschweiger danke ich für die konstruktive Zusammenarbeit beim DEL-Projekt. Bei den Mitgliedern der Nachwuchsgruppen, insbesondere Dr. Mateja Klika Škopić und Kathrin Jung bedanke ich mich für anregende Diskussionen und den Mitgliedern der AG Rauh, insbesondere Andreas Arndt, Bianca Wiedemann, Marina Keul, Dr. Julian Engel und Jörn Weisner danke ich für ihre Unterstützung bei verschiedenen Labormethoden. Ganz besonders bedanke ich mich bei der AG Koch (Anna Rudo, Dr. Mauro S. Nogueira, Dr. Jette Pretzel, Dr. Hitesh Patel, Christiane Ehrt, Tobias Brinkjost, Julia Jasper, Sarah Albert) für die stetige Unterstützung, die konstruktiven Ideen, Diskussionen und die vielen gemeinsamen Aktivitäten. Für ihren Einsatz für das BRD4-Projekt danke ich Dr. Jette Pretzel, Christiane Ehrt, Dr. Mauro S. Nogueira, Saskia Spitzer, Dr. Luis Bering und Anna Rudo. Dank geht ebenfalls an die ehemaligen Bachelor- und Masterstudenten: Sebastian Weigang, Saskia Spitzer, Robert Gehde und Dr. Luis Bering, welche mit ihrer Arbeit zu dieser Arbeit und auch zu den resultierenden Publikationen wichtige Beträge geleistet haben. Maßgeblich zum Gelingen dieser Arbeit beigetragen haben Mitglieder der AG Mutzel, insbesondere Till Schäfer, Andre Dorschinsky und Dr. Nils Kriege. Ich danke euch für die Zusammenarbeit, die vielen Diskussionen und die Einblicke in die Informatik, ohne die diese Arbeit nie hätte so interdisziplinär und spannend werden können. Für die Aufnahme der Kristalldaten bedanke ich mich bei den Mitgliedern der SLS-Trips. Mein Dank geht ebenfalls an das Sekretariat der AG Rauh (Martina Reibner, Petra Alhorn und Maria Sergani) für die Unterstützung, an die NMR- und AMD-Abteilung der CCB und an die Druckabteilung des MPIs (Melanie Wilkesmann und Claudia Pieczka) sowie an die DFG für die finanzielle Förderung im Rahmen des SPP 1736. Dr. Oliver Koch danke ich für die Überlassung des spannenden Themas.

Für die unermüdliche Unterstützung meiner Familie und meiner Freunde bedanke ich mich. Ohne sie wäre diese Arbeit nicht zu dem gereift, was sie ist.

Ich danke Ihnen und euch!

Eidesstattliche Versicherung (Affidavit)

Humbeck, Lina
Name, Vorname
(Surname, first name)

173609
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.

Dortmund, den 01.02.2019
Ort, Datum
(Place, date)

Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

Betrachtung der Ähnlichkeit von niedermolekularen Verbindungen unter Berücksichtigung der biologischen Aktivität

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.

Dortmund, den 01.02.2019
Ort, Datum
(Place, date)

Unterschrift
(Signature)