Prediction and Optimisation of Protein-Ligand Affinities by Integral Equation Theory

Dissertation zur Erlangung des Doktorgrades

Die Dissertation wurde im Zeitraum vom 09.2012 bis zum 05.2017 an der Fakultät für Chemie und Chemische Biologie der Technischen Universität Dortmund angefertigt.

M. Sc. Florian Mrugalla

Dortmund, 2017

Erstgutachter: Prof. Dr. Stefan M. Kast
Zweitgutachter: Prof. Dr. Holger Gohlke

Be nice.

John Niven, *The Second Coming*

Danksagung

Hiermit möchte ich mich bei Prof. Dr. Stefan Kast bedanken, der mir die Arbeit an diesem überaus spannenden Thema ermöglicht hat. Durch seine stetige Unterstützung bei gleichzeitiger Anregung zur Erweiterung des eigenen fachlichen Horizonts hat er meinen wissenschaftlichen Reifeprozess ungemein vorangetrieben.

An dieser Stelle sei natürlich auch Herrn Prof. Dr. Gohlke für die Übernahme des Zweitgutachtens gedankt.

Natürlich bedanke ich mich auch bei meiner gesamten Arbeitsgruppe für eine besonders angenehme, lustige und entspannte Zeit.

Besonders möchte ich mich bei meinen Kollegen Roland, Leonhard und Yannic bedanken, die mir immer mit Rat und Tat zur Seite standen und mit denen ich sehr viele lustige Stunden verlebt habe.

Ich möchte mich natürlich auch noch bei allen meinen Freunden bedanken die mich vor, während und nach dieser Arbeit begleitet haben bedanken. Hier sei Katja besonders hervorgehoben, da sie sich unermüdlich durch meine Bandwurmsätze gekämpft hat und somit großen Anteil am Lektorat dieser Arbeit hatte.

Unermesslicher Dank gilt meiner Familie, insbesondere meinen Eltern und meiner Schwester die mich jederzeit und in allen Lebenslagen unterstützen und ohne die ich garantiert heute nicht da wäre wo ich bin.

Mein größter Dank gilt Justina die immer ein offenes Ohr für mich hat, mich immer unterstützt, sogar über meine schlechten Witze lacht und diese Arbeit mittlerweile zu oft Korrektur gelesen hat.

Zusammenfassung

Das „three dimensional reference interaction site model“ (3D-RISM), hier insbesondere die Solvat-Solvat-Gleichung (*uu*), bietet Zugang zu atomweisen Beiträgen zum „potential of mean force“ (PMF). Das PMF setzt sich wiederum aus der direkten Wechselwirkung zwischen zwei Partnern und den durch die Solvatation vermittelten Beiträgen zusammen. Das PMF bietet zusätzlich Zugang zur freien Bindungsenthalpie, welche eine Schlüsselgröße für das Design neuer Moleküle in der Pharmazie ist.

Diese Arbeit beschäftigt sich hauptsächlich mit der Berechnung freier Bindungsenthalpien mit dem 3D-RISM-Ansatz und Methoden des maschinellen Lernens. Die abgedeckten Themengebiete reichen somit von den grundlegenden Prinzipien der Thermodynamik, repräsentiert durch den 3D-RISM-*uu*-Ansatz, bis hin zu empirischen Modellen basierend auf modernen Verfahren des maschinellen Lernens. Diese werden vertreten durch „deep neural networks“ und „boosted regression trees“.

Der erste Teil dieser Arbeit konzentriert sich auf die Vorstellung einer neuartigen Methode zur Bestimmung der Designrichtung im molekularen Raum. Dieses Werkzeug bezieht seine Information aus sogenannten „free energy derivatives“, welche relativ elegant und effizient innerhalb des 3D-RISM-*uu* Ansatzes definiert und berechnet werden können. Die dafür notwendigen theoretischen Grundlagen werden in dieser Arbeit gelegt und gleichzeitig wird eine Machbarkeitsstudie an dem gut charakterisierten 18-Krone-6-Ether-System durchgeführt. Diese Studie zeigt, dass sowohl experimentelle als auch theoretische Trends durch von 3D-RISM-*uu* berechnete PMFs und FEDs reproduziert werden können.

Diese aussichtsreichen Ergebnisse wurden zum Anlass genommen, diese Methode auf zwei Protein-Ligand-Systeme anzuwenden. Hierfür werden die entsprechenden Ligandenatome nacheinander entweder in die *apo*-Bindetasche oder die partiell belegte Bindetasche platziert. Beide Berechnungsmöglichkeiten liefern Zugang zu atomweisen PMFs und FEDs in Bezug auf typische Kraftfeldparameter. Zusätzlich wird auf die Stärken und Schwächen der gezeigten Methode eingegangen.

Im letzten Teil dieser Arbeit verlagert sich der Fokus darauf, eine neuartige „Scoring“-Funktion, welche auf struktureller Ligandeninformation beruht oder mit zusätzlichen atomweisen PMF-Werten berechnet durch 3D-RISM-*uu* zu „trainieren“. Für diesen „Trainingsprozess“ werden atomweise PMF-Werte mittels 3D-RISM-*uu* für eine Untermenge des „refined set“ und „core set“ der PDBbind-Datenbank berechnet. Dies kulminiert in einer „Scoring“-Funktion, die vergleichbare Ergebnisse zu anderen modernen „Scoring“-Funktionen liefert und bessere Ergebnisse in Bezug auf „klassische“ Scoring-Funktionen.

Abstract

The three dimensional reference interaction site model (3D RISM) in the form of the solute-solute (*uu*) equation allows one to calculate the atomwise contribution to the potential of mean force (PMF), which is composed of the direct interaction between two partners and solvation based contributions. The PMF is related to the binding free energy, which in turn is a key quantity for the design process of new molecular entities in pharmaceutical sciences.

This work revolves around the estimation of binding free energies with the 3D RISM and machine learning based methods. The range thus spans from fundamental thermodynamic principles represented by the 3D RISM-*uu* framework to empirical models based on modern machine learning, notably deep neural networks and boosted regression trees.

The first part of this work introduces a tool that could help to drive the design process in chemical space, which is highly desirable. This tool is based on free energy derivatives (FED), which can be easily defined and efficiently computed within the 3D RISM-*uu* framework, and which can provide a design direction that could ultimately lead to a better binder. The necessary theoretical basis is laid out in this work and tested in a proof of principle study on the well characterised 18-crown-6 ether system. In this study experimental and theoretical trends could be reproduced by PMFs and free energy derivatives calculated by 3D RISM-*uu*.

The promising results achieved in the aforementioned study were then applied to two protein ligand systems. For the protein ligand systems the respective ligand atoms are subsequently placed, either in the *apo* binding site or into the "partial *holo*" binding site that is made up of the supermolecule consisting of the protein and the partial ligand (ligand minus the atom in question). Both calculation schemes ultimately lead to atomwise information about the PMF and the respective FEDs with respect to typical non-bonded force field parameters. This study shows the possibilities and limitations of the aforementioned method.

For the last part of this work the focus shifts and it is demonstrated that it is possible to train a truly novel scoring function based on structural ligand information in the form of molecular fingerprints alone or in conjunction with atomwise PMF values calculated by 3D RISM-*uu*. For the training process atomwise PMF values were calculated for a subset of the PDBbind refined and core set. This culminated in scoring functions that are competitive with other modern machine learning based scoring functions and that outperform classical scoring functions significantly.

Table of contents

1 Introduction

1.1 Motivation

Designing new molecular entities and bringing them to the market is one of the grand challenges in chemistry and pharmaceutical sciences. The number of approved new molecular entities is stagnant in the last decade on a low level which was not seen in the flourishing years of drug development in the mid to late 20th century. This is despite the growing efforts made by the pharmaceutical industry by pouring more money into research and development.[1–3] This led to the incorporation of new approaches in the late 20th century, which were fuelled by advances in high-throughput screening and combinatorial chemistry.[4, 5] Nowadays, these are augmented by structure-based drug design, which has the premise that the activity of a ligand is encoded into the three dimensional structure[4, 6–8] and lead to the development of a plethora of docking and scoring functions.[8–13]

One of the hot topics, which garners the attention of the pharmaceutical industry is the role that weakly bound water plays in regard to binding thermodynamics.[14–19] This trend started in the 1990s and culminated into tools like SZMAP,[20] WaterFlap,[21] WaterRank,[22] WaterMap[23, 24] and others, which are used today and try to predict water binding sites. Although all these methods and tools help the medicinal chemist in the process from hit-to-lead design until to date rational design often boils down to a question of experience and so-called "chemical intuition."

Besides, the very important role that water plays in the binding process, the prediction of binding affinities is crucial during the first stages of the drug design process. So most of the questions that are asked during these stages boil down to the following two:

- Which molecule binds best to the target?
- Why does molecule X bind better to the target than molecule Y?

In order to answer the first question, theoretical chemistry and cheminformatics offer a plethora of methods that range from quantum mechanical calculations and molecular mechanics to the evaluation of empirical scoring functions. What the former two classes of methods offer in terms of accuracy they lack in terms of speed and vice versa for empirical scoring functions. Why do empirical scoring functions often do not offer a satisfactory amount of accuracy? And what is the missing piece of information for many of the aforementioned methods? A possible answer could be that a disruptive leap in either the translation of the underlying physics into a computationally tractable problem (better force fields) or the standardization and accurate measurement of the experimental database all empirical methods are relying on is needed. Because these breakthroughs are not in sight and clearly out of the scope of this work, another approach is to combine the best of both worlds. This means to design a model that is based on a relatively accurate description of the binding thermodynamics, including the crucial solvation contributions paired with relatively high computational efficiency (in the form of 3D RISM-*uu*[25–28]) and then leverage the predictive capabilities of modern machine learning methods to design a novel scoring function, thus compensating for noise and uncertainty.

Scoring functions are often categorized into four groups: force field-based, empirical, knowledge-based, and machine learning-based.[29, 30] Force field-based scoring functions basically rely on the calculation of the non-covalent interaction energy of the protein and the ligand in question and are augmented by the addition of solvation energy terms in form of continuum models like Poisson-Boltzmann[31, 32] (PB) or Generalized Born[33] (GB).[29] Scoring functions of this category benefit directly from advancements in the underlying force fields and representatives are for example `AutoDock`[34, 35] and `GOLD`[36]. Empirical scoring functions calculate the quality of a protein-ligand interaction through the weighted sum over rather arbitrary contributions. Frequent used descriptors are the number of rotatable bonds, the number of hydrogen bonds or the internal strain energies. The weights are determined by multivariate linear regression. The individual components can have a positive or negative effect on the resulting score.[29, 37] Unlike the aforementioned force field-based scoring functions, the functional form does not necessarily have a physical foundation.[29] Notable members of this class are `X-Score`[38] and `ChemScore`.[39] Knowledge-based scoring functions are based on the assumption that the protein ligand binding affinity can be described by the sum of all

pairwise interactions. These pairwise interactions are modelled as statistical potentials. DrugScore[40] and IT-Score[41, 42] are knowledge-based scoring functions. The fourth class of scoring functions combine a series of descriptors, which are fed into a machine learning method to derive binding affinity scores and are called machine learning-based scoring functions. These descriptors, or often called features, can consist of specific interactions: geometrical descriptors or ligand-based descriptors.[29] One significant difference of this type of scoring function from empirical scoring functions is the type of regression method, which is non-linear for machine learning-based scoring functions and linear for empirical scoring functions.[29, 37] In analogy, the two types of scoring functions share the need for a training set with experimentally determined binding affinities.[29] NNScore[43, 44] and RF-Score[45, 46] are representatives of this type of scoring function. According to Qurrat Ul Ain *et al.*,[37] one of the advantages of machine learning-based scoring functions is that they are not restricted to (multivariate) linear regression and a fixed functional form which is the case for "classical" scoring functions. This assumption is supported by studies where the performance of scoring functions could be improved by the substitution of linear regression through non-linear regression models.[37, 47–49] Neural networks and deep neural networks are also used as non-linear regression models. One example is NNScore[44] a shallow neural network with one hidden layer and 10 hidden neurons trained with the docking terms of Vina[50] 1.1.2 and features calculated by BINANA[51] on a handcrafted dataset based on the binding MOAD[52] and PDBbind.[53] More recently, Ashtawy *et al.*[54] combined either "bootstrap aggregation" (often called bagging) or "boosting"[55, 56] with a shallow neural network to enhance the predictive capabilities of their model. "Bagging" in the sense of machine learning means to combine an ensemble of trained models in an averaging manner. Boosting is a similar approach where the ensemble of models is combined through a weighted sum. The network architecture for both approaches consists of 20 hidden units for the hidden layer and as an input Ashtawy *et al.*[54] used a diverse set of descriptors that were extracted from various of-the-shelf scoring programs.[54] They trained their models on the refined set of the PDBbind[53] and reached Pearson correlation coefficients of $R = 0.80$ for the "bagging" and $R = 0.82$ for the boosting approach on their test set (core set of PDBbind). Thereby, they outperformed all other tested methods (including other non-linear regression models).[54] Wallach *et al.*[57] from Atomwise Inc. published an interesting paper on the "arXiv.org" server where they took a different route by using convolutional neural networks to classify active from inactive compounds. Their network architecture consisted of four 3D convolutional layers of varying filter sizes, followed by two fully connected hidden layers with

1024 hidden units in each layer.[57] More interesting than the network architecture is the design of the input representation: the input consisted of a cubic grid with 20 points in all directions and a spacing of 1 Å. Each grid cell contained structural information e.g. atom types. This 3D grid was unfolded into a 1D vector, which was then used to train the network. As input databases, mainly DUD-E[58] and a subset of the ChEMBL[59] database were used.[57] In the results they note that they achieve a "level of accuracy useful for drug discovery".[57] Deep neural networks can also be used for target prediction, which was done by Unterthiner *et al.*[60] for a subset of the ChEMBL[59] database. As input for their various tested methods and network architectures they used extended-connectivity fingerprints (ECFP).[61] The use of deep neural networks is not limited to the academic world. In 2012, Merck hosted a competition on kaggle[57, 60, 62, 63] (an online platform for data-science competitions) with the goal of testing the performance of modern machine learning methods on QSAR problems. The winning team made heavy use of a multi task deep neural network and was able to achieve a relative improvement of 15 % over the in-house baseline models of Merck.[62, 63]

As far as the author knows, no scoring function, so far, was trained on thermodynamic data calculated by 3D RISM-*uu*. Nonetheless, a little synopsis of the role RISM has in molecular modelling is given here. For example Genheden *et al.*[64] approximated the binding free energy of protein ligand complexes through sampling of the conformational degrees of freedom with MD simulation coupled with 3D RISM-*uv* calculations (called MM-3D-RISM-KH[65]) of simulation snapshots in a MM/PBSA(GBSA) manner.[64] Imai *et al.*[66] took a different route: they used mixtures of water and a drug fragment as a solvent and calculate the respective pair distribution functions of all components and the protein. With these solvent site distributions they tried to detect potential binding sites on the protein surface and also to deduce possible binding modes of the fragment in the active site.[66] Nikolić *et al.*[67] picked up this idea and implemented a new docking approach based on the PMF calculated by 3D RISM-*uv* into `AutoDock`,[34, 35] which they call 3D-RISM-DOCK. The only application of 3D RISM-*uu* so far in the field of molecular modelling or structure based drug design is a proof of principle study by Kiyota[68] *et al.* where the aim was to reproduce the binding mode of a model system consisting of aspirin and phospholipase A2.[68] But applications of 3D RISM and in particular the solute-solute equations (*uu*) to the prediction of binding affinities or molecular modelling in general is rather limited.[64, 66, 68, 69]

Returning to the second question which is even more complex and somewhat of a holy grail in the medicinal chemistry community. Subtle changes in the ligand chemistry can have a huge impact on the measured or calculated binding affinity. Most often the decision, where and which derivatisation to make, is driven by chemical intuition or empirical rules like the rule of five.[70, 71]

As laid out in Ref. [72], designing functional molecular systems essentially means the process of translating desired properties of a material or a biologically active substance into chemical structure. Since there does not exist a one-to-one mapping between a desired (continuous) property, such as a specific band gap, elastic constant, or protein-ligand binding affinity, and chemical structure space, molecular design can be rational only to a limited extent. These limits are defined, on one hand, by the discreteness of chemical structure space (not every conceivable or desirable value of a property can be realized chemically) and, on the other hand directly related to the first issue, by knowledge accumulated in the past, namely the measured or theoretically predicted properties of given chemical compounds. Moreover, the space of potentially useful structures is huge (for pharmacologically relevant compounds, number estimates range from 10^{23} to 10^{60} [73]), giving rise to the opportunity that several, even completely dissimilar chemistries can have properties close to the desired value. Designing molecules is therefore characterized by an underdetermined, inverse problem subject to additional constraints such as synthetic accessibility, minimization of unwanted side effects, as well as economic and legal factors like minimizing production costs or maximizing likelihood of patentability and premarket approval.

Focusing now on pharmacological problems, as already mentioned above, the pharmaceutical industry is facing a dire problem related to the fundamental design issue. Despite about $50 billion spent annually on research and development only 20 new drugs are released per year.[74] One of the reasons for the ever growing gap between costs and return is the fact that only about 3% of the initiated drug discovery projects make it to a marketable drug.[75] One reason for the failure of this daunting procedure is related to a very early stage of development, the so-called lead-optimization phase, where the problem is to decide where and how to modify a ligand molecule in order to get a more favourable binding to a target protein. The property or key thermodynamic quantity defining the design goal in this case is the (standard) free energy of binding (ΔG_{bind}, omitting the "standard" superscript for simplicity). The broad spectrum of methods for calculating these free energies range from docking algorithms based on empirical

scoring functions (as described in detail above), which can be evaluated comparatively fast, to explicit-solvent fully atomic molecular dynamics (MD) simulations, see Refs. [76-78] for recent overviews. The latter, presently representing the method with highest level of physical detail achievable, requires orders of magnitude more computing time than the former simpler, though far less accurate techniques.[1]

Yet, even though substantial advances have been and are currently being made in the field of predicting protein-ligand thermodynamics, such methods, frequently combined with virtual screening techniques to reduce the chemical search space,[76] do not directly address the primary design goal but provide posteriori data only to be fed back into an iterative design cycle. Clearly, progress can be made by defining a search direction in property space, which in this case would be equivalent to define a (binding) free energy derivative (FED) with respect to certain parameters that define variations of chemical space. The simplest way to this end is to vary protein-ligand (and therefore simultaneously ligand-solvent) interaction parameters taken from model potentials. For instance, locally changing a site charge and/or apolar atomic size/interaction strength parameters can be viewed as virtual substitutions on otherwise unchanged scaffolds, for which a derivative can be mathematically defined. Such a concept of deriving a FED based on free energy MD simulations was proposed more than two decades ago by several authors[77–79] and has been further explored with more or less promising results.[82-84] In a related approach, van Gunsteren and co-workers have devised methods to compute free energy changes simultaneously for several target states (representing different chemistries) from simulation of only a single reference system.[83, 84] The drawback of these techniques which led to limited acceptance in practical applications is certainly the high computational demand involved with such MD simulations. As an alternative, much faster to evaluate yet more approximate models for "charge optimization" have been proposed based on a minimization of electrostatic energy within dielectric continuum solvation theory.[85–87] This method turned out to be rather insensitive to changes in the ligand conformation and provides reasonable results with rigid ligands.[88, 89]

To make further progress, it is desirable to combine the level of physical detail of explicit MD methods with the computational efficiency of implicit methods such as continuum models. In this work an alternative route to binding FEDs is introduced on the basis of liquid state

[1] Reused in part with permissions from F. Mrugalla, S. M. Kast, *J. Phys.: Cond. Matter* **2016**, *28*, 344004. © 2016 IOP Publishing Ltd.

theory in the form of the 3D RISM[90–93] which, as a primary result, yields approximate solute-solvent (*uv*) molecule-solvent site distribution functions on a 3D grid from which thermodynamic quantities, including the solvation free energy, can be derived analytically within certain approximations. Unlike continuum solvation models 3D RISM-*uv* theory is capable of retaining the directionality of solute(*u*)-solvent(*v*) interactions based on the same interaction potential that could be used in MD simulations, thereby retaining the atomic level of detail. While the *uv* formulation needs properties of the pure solvent as input (the solvent site-site susceptibility χ derived from solution to a simpler 1D RISM-*vv* equation or taken from MD simulations), it is also possible to extend the hierarchy toward an integral equation between two infinitely diluted solute species, the solute-solute (*uu*) equation.[25, 26, 28, 94] The particular appeal of the *uu* theory is related to the fact that it yields the so-called potential of mean force (PMF), *i.e.* the free energy surface governing complex formation analytically and non-iteratively starting from precomputed *uv* solutions for the individual partners only. Derivatives of this quantity with respect to interaction potential parameters therefore serve the goal to define a possible design direction on the basis of a physically detailed yet compared to MD orders of magnitude more efficiently computable theory.

1.2 Aims of this work

The aim of this work is, to establish the use of the 3D reference interaction site model,[94, 95, 97, 98] specifically the solute-solute equation (3D RISM-*uu*),[25–28] for drug design purposes. This aim is pursued by two different means: First the theoretical groundwork has to be established and transferred into a working numerical implementation. This is followed by the proposition of several new ideas regarding the use of 3D RISM-*uu* within the drug design process. These ideas culminate into several proof of principle studies in which the weaknesses and strengths of each of the approaches are investigated.

In detail this means that after introducing the necessary theoretical basics, which is done in the next chapter, the focus of this work shifts to the introduction of a novel approach to generate design directions for a given molecular system, namely the combination of 3D RISM-*uu* and free energy derivatives in the spirit of the work done by others on molecular dynamics simulation.[77–79] This study is done on the extensively described 18-crown-6 ether system and

shows that the PMF topography calculated by 3D RISM-*uu* is in qualitatively good agreement compared to topography calculated by 3D RISM-*uv* or thermodynamic integration. The sufficient agreement of the PMF topography permits to calculate free energy derivatives, which yield meaningful results regarding the optimal binding partner for the mentioned crown ether system.

This successful application of free energy derivatives sparked further interest into the application on protein ligand systems. To apply free energy derivatives in this setting, two different calculation schemes were devised (later called *apo* and partial *holo* scheme), implemented and tested. The concrete protein ligand system under scrutiny is: tRNA guanine transglycosylase (TGT) bound to two different aminoquinazolin derivates, that have the interesting property of being matched molecular pairs.[97] They also show a rather distinct difference in their binding affinity with only subtle changes in the binding mode, making them an almost optimal example to test the design tool proposed earlier. Free energy derivatives are calculated on an atomwise basis with respect to the non-bonded force field parameters and analysed to elucidate the most likely cause for the difference in binding affinity and also show the limitations of the approach.

The last chapter of this work aims towards the generation of a new type of scoring function based on the aforementioned PMFs calculated by 3D RISM-*uu* and modern machine learning techniques. A workflow for automated parametrisation and calculation of 3D RISM-*uu* PMFs is devised and applied to a subset of the PDBbind.[53] After that different compositions of feature sets and the underlying experimental data are used to train several scoring functions with either deep neural networks or boosted regression trees.[98, 99]

2 Theoretical background

2.1 Host-Guest binding

In the most general sense, a binding process between a host system and a guest (here often interchangeably used as protein-ligand) can be described through the binding constant (interchangeably used as association constant) which can be defined as:

$$K_b = \frac{[LP]}{[L][P]}. \tag{1}$$

Here [LP], [L], and [P] are the respective equilibrium concentrations of the bound and unbound state. Following this equation the standard binding free energy can be defined as:

$$\Delta G_b^\circ \equiv \mu_{sol,LP}^\circ - \mu_{sol,L}^\circ - \mu_{sol,P}^\circ = -k_B T \ln\left(C^\circ K_b\right) \equiv \Delta G, \tag{2}$$

where k_b is the Boltzmann constant, T is the temperature, C° is the standard concentration of 1 mol/l, and μ_{sol}° is the chemical potential of the respective species in solution. Calculating standard binding free energies is an arduous procedure.[100–103] Therefore often only the relative free energy change between to species is calculated which can be written as:

$$\Delta\Delta G_{L_1P \to L_2P} = \Delta G_{b,L_2P} - \Delta G_{b,L_1P}, \tag{3}$$

which is a computationally better tractable problem. If the binding constants of two complexes L_1P and L_2P are known the relative free energy change is given as:[103]

$$\Delta\Delta G_{L_1P \to L_2P} = k_B T \ln\left(\frac{K_{b,L_2P}}{K_{b,L_1P}}\right). \tag{4}$$

Relative free energy changes are computationally accessible through various methods, for example thermodynamic integration coupled to molecular dynamics, which is described in chapter 2.4.1

Experimentally binding affinities can also be measured as dissociation constants which are defined as

$$K_{\mathrm{d}} = \frac{1}{K_{\mathrm{b}}}. \tag{5}$$

Often the half maximal inhibition constant, called IC_{50} or the inhibition constant K_i is measured. The interconversion between IC_{50} values and K_i is possible with the Cheng-Prusoff equation, which states for reactions where one ligand is involved:

$$\mathrm{IC}_{50} = K_{\mathrm{i}}\left(1 + \frac{[\mathrm{S}]}{K_{\mathrm{m}}}\right), \tag{6}$$

where K_m is the Michaelis constant, which is a kinetic constant, and [S] is the substrate concentration.[104] Another possible way to the determination of K_i values is through the following relationship:

$$\frac{\mathrm{IC}_{50,1}}{\mathrm{IC}_{50,2}} = \frac{K_{\mathrm{i},1}}{K_{\mathrm{i},2}}, \tag{7}$$

where only the IC_{50} values of all involved species (1 and 2 in this case) and the K_i value for one species, either 1 or 2 are necessary.

On a molecular level the direct interactions that most commonly occur are hydrogen bonds, halogen bonds, interactions between halogens and aromatic rings, hydrophobic interactions, and π-π-interactions of aromatic ring systems.[105–108] The energetic contributions of each of these interactions can vary. The indirect interactions on the other hand are governed by the solvent e.g. water and can have a greater effect on the resulting binding free energy than the direct interactions.[16, 109–113] Indirect interactions can be traced back to both enthalpic and entropic contributions, which can arise through a direct water network bridging the ligand to the protein, or through entropic contributions because of wetting/dewetting events upon binding.[16, 17, 111, 114, 115]

2.2 Reference interaction site model (RISM)

Classical density functional theory facilitates the connection between particle densities to the free energy of a fluid system in thermodynamic equilibrium[116, 117]. The local particle densities $\rho_\gamma(r)$ of particle type γ are connected to the pair distribution function $g(r)$ between particles γ and α through:

$$\rho_\gamma(r) = \rho_\gamma g_\gamma(r). \tag{8}$$

Here ρ_γ represents the bulk site density of the system. The pair distribution function $g_\gamma(r)$ (see Figure 1) describes the interaction between two particles and is normalised to the bulk

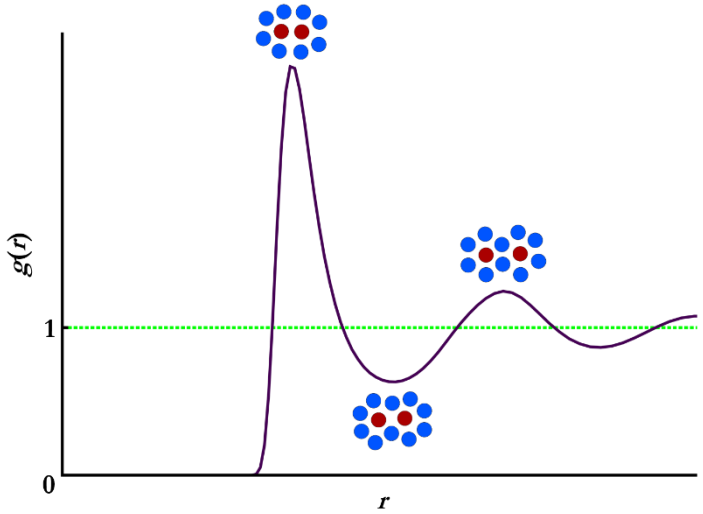

Figure 1: Idealised illustration of a radial pair distribution function $g(r)$. Areas of favourable interaction have $g > 1$ and areas of unfavourable interaction have $g < 1$.

density of the system in question. The pair distribution function is also one of the key quantities in this work and relates to the total correlation function $h_\gamma(r)$ in the following way:[116]

$$h(r) = g(r) - 1. \tag{9}$$

Through ground breaking work done by Leonard Ornstein and Frederik Zernike in the early 20th century that culminated in the Ornstein-Zernike equation, which has the form:[118]

$$h(r) = c(r) + \rho \int c(r)h(r)\mathrm{d}r, \tag{10}$$

where h stands for the total correlation function, c for the direct correlation and ρ for the particle density. The total correlation function $h(r)$, which oneself is interested in, describes all solvent interactions between the particles. The direct correlation function, the first term on the right hand side of equation (10), accounts for the direct solvent mediated interaction between the particles. The second term of equation (10) in the form of the integral describes all the indirect interactions solvent mediated interactions (see Figure 2 for a sketch). It is easy to see that equation (10) cannot be solved analytically, and a second equation is needed to close the system of equations. This equation is known as closure and takes the general form of:

$$h(r) = \exp\left[\beta U(r) + \rho \int c(r)h(r)\mathrm{d}r + B\right] - 1. \qquad (11)$$

In the closure expression $u(r)$ represents a pairwise additive potential and $\beta \equiv 1/k_b$ is the inverse thermodynamic temperature. The so called bridge function B describes higher than second order correlations and is not analytically accessible. Numerically, the bridge function can be approximated, but a closed form does not exist.[116, 119–121]

So far only the pure atomic case was considered, which clearly has limited applicability in real world scenarios. Consequently, the molecular Ornstein-Zernike equation was derived, which is applicable to molecular problems and has the form:[116, 122]

$$h(\mathbf{r}_{12}, \mathbf{\Omega}_1, \mathbf{\Omega}_2) = c(\mathbf{r}_{12}, \mathbf{\Omega}_1, \mathbf{\Omega}_2) + \frac{\rho}{8\pi^2} \int\int c(\mathbf{r}_{13}, \mathbf{\Omega}_1, \mathbf{\Omega}_3)h(\mathbf{r}_{32}, \mathbf{\Omega}_3, \mathbf{\Omega}_2)\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{\Omega}. \qquad (12)$$

In contrast to the Ornstein Zernike equation, the relative orientation between two molecules is added in the form of Euler angles $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$. While analytically exact, finding numerical solutions to the six dimensional molecular Ornstein Zernike equation is cumbersome.[123, 124] This limits the applicability and is rooted in the high dimensionality and integration with regard to the Euler angles.

To avoid these problems, Chandler and Anderson went back several levels of dimensionality and derived the one dimensional reference interaction site model (1D RISM), often also called the site-site Ornstein Zernike equation.[90, 91] Conceptually a molecule is broken down into its interaction sites, for example water is broken down into the three sites O, H and H, which can reduced into O and H through symmetrical considerations. The 1D RISM equation can be written in "matrix form:"

$$\mathbf{h} = \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\omega} + \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\rho}\mathbf{h}. \tag{13}$$

In equation (13) the matrices for the already mentioned total and direct correlation functions (**h** and **c**) and the density matrix **ρ** (for the infinite dilution case) are constructed in the following way:

$$\mathbf{h} = \begin{pmatrix} \mathbf{h}^{vv} & \mathbf{h}^{vu} \\ \mathbf{h}^{uv} & \mathbf{h}^{uu} \end{pmatrix}, \mathbf{c} = \begin{pmatrix} \mathbf{c}^{vv} & \mathbf{c}^{vu} \\ \mathbf{c}^{uv} & \mathbf{c}^{uu} \end{pmatrix}, \boldsymbol{\rho} = \begin{pmatrix} \rho^{v} & 0 \\ 0 & 0 \end{pmatrix}. \tag{14}$$

Here $\mathbf{h}^{uv}$ equals $\left(h_{\alpha\gamma}(r_{\alpha\gamma})\right)^{uv}$ which means that the total correlation function consists of all sites α of the solute u and all sites γ of the solvent v. The matrix $\mathbf{c}^{uv}$ is constructed in the same way as $\mathbf{h}^{uv}$. In principle α and γ can consist of an arbitrary number of atoms which are represented by their position in relation to the respective nuclei or (in 3D RISM) only solvent sites are decomposed. Based on the water example from above this means that a 1D RISM solution of one water molecule (u) in water (v) actually consists of three total correlation functions: h^{OO}, h^{OH} and h^{HH}. The reduction to a site-site model would lead to a loss of all information about the intramolecular structure of the molecule in question due to the rotational average

$$c(r) = \sum_{\alpha}\sum_{\gamma} c_{\alpha\gamma}(r_{\alpha\gamma}). \tag{15}$$

To circumvent this loss of information a new function in the form of,

$$\omega(r_{\alpha\gamma}) = \frac{\delta(|r_{\alpha\gamma} - l_{\alpha\gamma}|)}{4\pi l_{\alpha\gamma}^{\,2}}, \tag{16}$$

is introduced. The intramolecular distances are encoded in $l_{\alpha\gamma}$ and $\delta(x)$ is the Dirac delta function. As can be easily seen by some linear algebra,[25] the expansion of equation (13) leads to three distinct equations, namely the solvent-solvent (vv), solute-solvent (*uv*) and solute-solute (*uu*) equations:

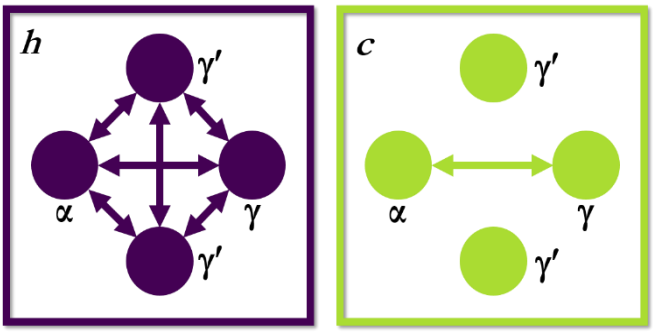

Figure 2: Schematic illustration of the interactions that are governed by the total (*h*) or direct (*c*) correlation functions.

$$\mathbf{h}^{vv} = \boldsymbol{\omega}^{v} * \mathbf{c}^{vv} * \boldsymbol{\omega}^{v} + \boldsymbol{\omega}^{v} * \mathbf{c}^{vv} * \boldsymbol{\rho}^{v}\mathbf{h}^{vv}, \tag{17}$$

$$\mathbf{h}^{uv} = \boldsymbol{\omega}^{u} * \mathbf{c}^{uv} * \boldsymbol{\omega}^{v} + \boldsymbol{\omega}^{u} * \mathbf{c}^{uv} * \boldsymbol{\rho}^{v}\mathbf{h}^{vv}, \tag{18}$$

$$\mathbf{h}^{uu} = \boldsymbol{\omega}^{u} * \mathbf{c}^{uu} * \boldsymbol{\omega}^{u} + \boldsymbol{\omega}^{u} * \mathbf{c}^{uv} * \boldsymbol{\rho}^{v}\mathbf{h}^{vu}. \tag{19}$$

These three equations have to be solved in consecutive order starting with the iterative solution of the 1D RISM-*vv* equation; this solution is then needed to calculate an iterative solution of the 1D RISM-*uv* equation. As a matter of fact, the solution to equation (19) can be calculated non-iteratively, which is a huge advantage of the 1D RISM-*uu* equation and will play a crucial role in the following work.

In the 1990s an expansion of the 1D RISM equation to three dimensions was derived by several groups[92, 93, 95, 96] and yields, for the solute-solvent case,

$$h_{\gamma}^{uv}(\mathbf{r}) = \sum_{\gamma'} \rho_{\gamma'}^{-1} \int c_{\gamma'}^{uv}(\mathbf{r}-\mathbf{r}')\chi_{\gamma\gamma'}(\mathbf{r}_{\gamma'\gamma})\mathrm{d}\mathbf{r}_{\gamma'}. \tag{20}$$

To solve equation (20) two requirements have to be met: (1) the so called solvent-susceptibility, which is encoded in the $\chi_{\gamma\gamma'}$ function, has to be precalculated with 1D RISM-*vv*. (2) In the same manner as in equation (11) a closure relation is needed. Commonly, the hypernetted-chain-closure (HNC) is used:[125, 126]

$$h_{\gamma}^{uv}(\mathbf{r}) = \exp\left[t_{\gamma}^{uv,R}(\mathbf{r})\right] - 1 \tag{21}$$

where the bridge function *B* from equation (11) is set equal to zero and the new function $t_{\gamma}^{uv,R}$ is introduced, which is defined as

$$t_{\gamma}^{uv,R}(\mathbf{r}) = t_{\gamma}^{uv,R}(\mathbf{r}) - \beta U_{\gamma}(\mathbf{r}) = h_{\gamma}^{uv}(\mathbf{r}) - c_{\gamma}^{uv}(\mathbf{r}) - \beta U_{\gamma}(\mathbf{r}) \tag{22}$$

and can be interpreted as a "renormalized" indirect correlation function and $U_{\gamma}(\mathbf{r})$ is the pairwise additive potential between all interaction sites of the solute and the solvent site γ. The pairwise potential is usually comprised of

$$U_{\gamma,\mathrm{Coulomb}}(\mathbf{r}) = \sum_{\alpha} \frac{q_{\alpha}q_{\gamma}}{4\pi\varepsilon_{0}|\mathbf{r}-\mathbf{r}_{\alpha}|}, \tag{23}$$

$$U_{\gamma,\mathrm{Dispersion,Rep}}(\mathbf{r}) = \sum_{\alpha} 4\varepsilon_{\alpha\gamma}\left(\left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r}-\mathbf{r}_{\alpha}|}\right)^{12} - \left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r}-\mathbf{r}_{\alpha}|}\right)^{6}\right) \tag{24}$$

where the electrostatic interaction is modelled by the Coulomb potential with q being the partial charge of particle type α or γ, ε_0 as the dielectric permittivity of the vacuum and the distance $|\mathbf{r}-\mathbf{r}_\alpha|$ between the solute sites α and solvent sites γ. The dispersion interaction is represented in the form of the Lennard-Jones potential where $\varepsilon_{\alpha\gamma}$ corresponds to the "well depth" of the potential and $\sigma_{\alpha\gamma}$ can be interpreted as the contact distance. Both parameters ε and σ are commonly taken from molecular mechanics force fields. One drawback of the hypernetted chain closure is the fact that numerical convergence of the equation system cannot be guaranteed and getting the equation system to convergence can sometimes be called a "black art," or is impossible, respectively. In light of this, another set of closures developed in the group of Kast *et al.*[127] based on the partial series expansion of order k (PSE-k) of the HNC closure is often used and have the form:

$$h_\gamma^{uv}(\mathbf{r}) = \begin{cases} \sum_{i=0}^{k} \left(t_\gamma^{uv,R}(\mathbf{r})\right)^i / i! - 1 & \Leftrightarrow \quad t_\gamma^{uv,R}(\mathbf{r}) > 0 \\ \exp\left[t_\gamma^{uv,R}(\mathbf{r})\right] - 1 & \Leftrightarrow \quad t_\gamma^{uv,R}(\mathbf{r}) \le 0 \end{cases} . \tag{25}$$

These PSE-k closures approximate the HNC closure and show a good-tempered convergence behaviour and are therefore often preferred over the HNC closure.[28, 127] In the literature the PSE closure of order one is often called the Kovalenko-Hirata (KH) closure.[27, 94] For these closures a closed form of the chemical excess potential exists and can be defined as

$$\mu^{ex} = -\beta^{-1} \sum_\gamma \rho_\gamma \int \left[\frac{1}{2}\left(h_\gamma^{uv}\right)^2(\mathbf{r}) - \frac{1}{2} h_\gamma^{uv}(\mathbf{r}) c_\gamma^{uv}(\mathbf{r}) - c_\gamma^{uv}(\mathbf{r}) - \Theta\left(h_\gamma^{uv}(\mathbf{r})\right) \frac{\left[t_\gamma^{uv,R}(\mathbf{r})\right]^{n+1}}{(n+1)!} \right] \tag{26}$$

where Θ is the Heaviside step function which vanishes in the case of the HNC closure. Equation (2) and equation (26) show the connection between 3D RISM and the free energy.

The generalisation of the 3D RISM equation for the solute-solute (*uu*) case[25–28] can be written as

$$\begin{aligned} h^{uu}(\mathbf{R}_{12}, \Omega_{12}) &= c^{uu}(\mathbf{R}_{12}, \Omega_{12}) + \sum_\gamma c_{1\gamma}^{uv} * \rho_\gamma h_{2\gamma}^{uv}(\mathbf{R}_{12}, \Omega_{12}) \\ &\equiv c^{uu}(\mathbf{R}_{12}, \Omega_{12}) + \eta(\mathbf{R}_{12}, \Omega_{12}) \end{aligned} \tag{27}$$

where the indices 1 and 2 represent two different solute species in question. For 3D RISM-*uu* calculations the *uv* results of the two partners 1 and 2 are needed. Because the *uv* results are calculated for fixed orientations, the 3D RISM-*uu* equation formally depends not only on the coordinates of the two solutes but additionally on the relative orientation of the two partners.

The HNC closure in the 3D RISM-*uu* case can be written in the same form as for 3D RISM-*uv*, which is shown in equation (21).

2.2.1 The potential of mean force and derived quantities in the case of 3D RISM

The potential of mean force is a fundamental quantity in chemistry and describes the free energy change along a reaction pathway[128] in solution. It can be defined as

$$w(\mathbf{r}) = U(\mathbf{r}) + G^{\text{solv}}(\mathbf{r}) = U(\mathbf{r}) + \Delta\mu^{\text{ex}} \tag{28}$$

where $U(\mathbf{r})$ is the direct potential of the molecule in question and $G^{\text{solv}}(\mathbf{r})$ is the free energy contribution of the solvation process, which equals $\Delta\mu^{\text{ex}}$ in the case of unpolarisable and rigid molecules (see Figure 3 for visualisation). The PMF can be also linked to the pair distribution function with the reversible work theorem[116] which states that:

$$w(\mathbf{r}) = -\beta^{-1} \ln g(\mathbf{r}). \tag{29}$$

In the specific case of a complex consisting of two solute molecules (*uu*) the PMF in the 3D RISM-*uu* case can be derived by restructuring equation (27) and equation (29) and the *uu* analogue of the HNC closure into,[27, 28, 129]

$$w^{uu}(\mathbf{R}_{12}, \mathbf{\Omega}_{12}) = U_{12}(\mathbf{R}_{12}, \mathbf{\Omega}_{12}) - \beta^{-1}\eta(\mathbf{R}_{12}, \mathbf{\Omega}_{12}) \tag{30}$$

which is directly solvable. The vacuum interaction potential $U_{12}(\mathbf{R}_{12}, \mathbf{\Omega}_{12})$ is the sum of all pairwise contributions as seen in equations (23) and (24), the function $\eta(\mathbf{R}_{12}, \mathbf{\Omega}_{12})$ can be interpreted as the solvent mediated influence on the resulting PMF (in this work this quantity is often abbreviated as w^{γ} after multiplication with β^{1}). If one solute species *u* is spherically symmetric, the orientational dependence can be dropped without loss of generality. This leads to direct access to the PMF, which is only dependent on the coordinates of both solute species. Because the interaction potential U_{12} as well as η are long ranged functions, a renormalisation procedure is necessary (see below).[28] In principle it is also possible to calculate ΔG_{bind} from w^{uu}, due to the PMF being a difference of state function, by integration of the bound region, which has to be defined beforehand.[103]

It is also possible to gain access to the explicit PMF of two rigid solute species with subsequent 3D RISM-*uv* calculations by

$$w^{\mathrm{expl}}(\mathbf{R}_{12},\mathbf{\Omega}_{12}) = \mu^{\mathrm{ex}}_{\mathrm{complex}}(\mathbf{R}_{12},\mathbf{\Omega}_{12}) - \mu^{\mathrm{ex}}_{1} - \mu^{\mathrm{ex}}_{2} + U_{12}(\mathbf{R}_{12},\mathbf{\Omega}_{12}). \tag{31}$$

Here μ^{ex} as defined in equation (26) has to be calculated for the complex consisting of the two solute species in question and the separate solute species (to account for the reference state). After that the PMF is obtained by adding the vacuum interaction potential.

Because the PMF can be computed directly and with high computational efficiency with 3D RISM-*uu*, it is possible to derive so called "free energy derivatives" on their basis. Free energy derivatives (FEDs) were first described by Pearlmen in the 1990s[79] and, following their narrative, they can be defined for 3D RISM-*uu* as

$$\frac{\partial w^{uu}(\mathbf{R}_{12},\mathbf{\Omega}_{12})}{\partial \varphi_i} \tag{32}$$

where φ_i acts as a substitute for one of the parameters (ε, σ, q) in the interaction potential between the two solutes. This interaction potential can have the following form:

$$U_{12}(|\mathbf{r}_1 - \mathbf{r}_2|) \equiv U_{12}(r_{12}) = 4\varepsilon_{12}\left[\left(\frac{\sigma_{12}}{r_{12}}\right)^{12} - \left(\frac{\sigma_{12}}{r_{12}}\right)^{6}\right] + \frac{q_1 q_2}{4\pi\varepsilon_0 r_{12}}. \tag{33}$$

where ε_{12} is a parameter of the Lennard-Jones potential that describes the well depth, σ_{12} is the point of the zero-crossing and q_1/q_2 are the partial charges of molecules 1 and 2.

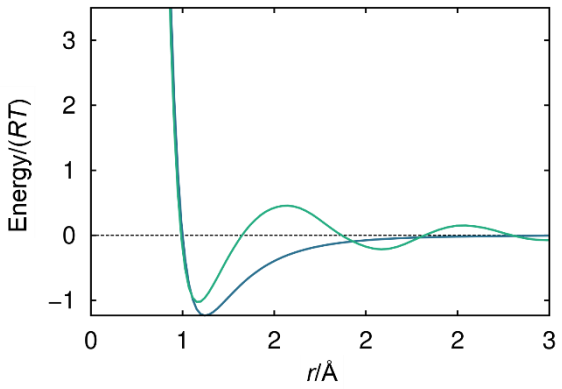

Figure 3: Schematic graph of the direct vacuum potential (blue) and the PMF (green). Here energy is either the direct interaction energy between two species (blue) or the interaction energy in solution (green).

2.3 Methodology of the renormalisation of long-range interactions

This chapter deals with the technical side of the necessary renormalisation procedure that is formaly laid out in Ref. [28]. As shown in equations (27) and (30) the PMF for the 3D RISM-*uu* case can be formulated in the following way:

$$w^{uu}(\mathbf{R}_{12},\mathbf{\Omega}_{12}) = U_{12}(\mathbf{R}_{12},\mathbf{\Omega}_{12}) - \beta^{-1}\eta(\mathbf{R}_{12},\mathbf{\Omega}_{12}). \tag{34}$$

A renormalized potential $\overline{U}_{12}^{S}(\mathbf{r})$ for dielectric solvents, to which this work is restricted, can be derived from:[28]

$$\beta U_{12}^{L}(\mathbf{r}) - \eta_{12}(\mathbf{r}) \sim \frac{1}{\varepsilon}\beta U_{12}^{L}(|r| \to \infty), \tag{35}$$

where $U_{12}^{L}(\mathbf{r})$ is the long range potential, whose computation has to be avoided, and ε is the dielectric constant. This leads to:[28]

$$\overline{U}_{12}^{S}(\mathbf{r}) = U_{12}^{S}(\mathbf{r}) + \frac{1}{\varepsilon}U_{12}^{L}(\mathbf{r}) \tag{36}$$

where $U_{12}^{S}(\mathbf{r})$ corresponds to the full real-space intermolecular potential (Lennard-Jones and Coulomb) and $1/\varepsilon U_{12}^{L}(\mathbf{r})$ is the weighted long range part.[28] A second partitioning of the potential is also needed, which leads to

$$\overline{U}_{12}^{L}(\mathbf{r}) = \frac{\varepsilon - 1}{\varepsilon}U_{12}^{L}(\mathbf{r}). \tag{37}$$

Equations (36) and (37) are computed in a straightforward way in the case of a dielectric solvent and in the case of an electrolyte they are replaced with their unweighted counterparts.[28]

The renormalized $\overline{\eta}_{12}^{s}(\mathbf{r})$ function as defined in Ref. [28] has the form

$$\begin{aligned}\overline{\eta}_{12}^{s}(\mathbf{r}) = &-\beta\left(\overline{U}_{12}^{L}(\mathbf{r}) - \overline{U}_{12}^{L(0)}(r)\right)\\ &+\sum_{\gamma}\left(c_{1\gamma}^{uv} + \beta\overline{U}_{1\gamma}^{L(0)}\right) * \rho_{\gamma}h_{2\gamma}^{uv}(\mathbf{r})\\ &-\overline{U}_{12}^{L(0)}(r)\\ &-\sum_{\gamma}\beta\overline{U}_{1\gamma}^{L(0)} * \rho_{\gamma}h_{2\gamma}^{uv}(\mathbf{r})\end{aligned} \tag{38}$$

Here $\overline{U}_{12}^{L(0)}(\mathbf{r})$ is the monopole potential which can be written as

$$\overline{U}_{12}^{L}(r) = \mathrm{erf}(k|\mathbf{r}-\mathbf{R}_0|)\frac{q_1\sum_2 q_2}{\mathbf{r}-\mathbf{R_0}}. \quad (39)$$

To ensure correct treatment the four terms of equation (38) have to be evaluated in a distinct order and manner.

The first term, $\overline{U}_{12}^{L}(\mathbf{r})-\overline{U}_{12}^{L(0)}(\mathbf{r})$, is evaluated on a 3D grid to acquire $\overline{U}_{12}^{L}$ and $\overline{U}_{12}^{L(0)}$. The second term is evaluated on 3D grids which implies that $\beta\overline{U}_{1\gamma}^{L(0)}$ is calculated in *k*-space and added to $c_{1\gamma}^{mw}$, subsequently the *k*(0) element of the 1D $h_{2\gamma}^{mw}$ function is extrapolated, as described in Ref. [130], and the result is interpolated onto a 3D grid. At last the convolution product of the second term is calculated. Term 2 is subtracted from Term 1 and the resulting *k*-space function is transformed with the reverse 3D FFT.[131, 132]

The third term, $\overline{U}_{12}^{L(0)}(r)$, is evaluated analytically on the 1D grid. The monopole potential $\beta\overline{U}_{1\gamma}^{L(0)}$ of the fourth term is evaluated in 1D-*k*-space analytically,[28] which is then followed by the convolution product. After that term three and four are added in 1D-*k*-space, followed by reverse 1D FFT[131, 132] and interpolation onto the real space 3D grid.

In the last step the two resulting 3D real space grids are added and yield $\overline{\eta}_{12}^{s}$. Pseudocode for the calculation of the renormalized η-function can be found in the appendix of this work.

2.4 Molecular dynamics (MD) simulation

MD simulations have come a long way from the 3 ps long trajectories (864 argon atoms)[133] of the early 1960s to millisecond long calculations (~17000 atoms)[134] done on modern super computers and special purpose hardware. In this time frame they evolved from a niche to the scientific mainstream.

Regardless of this evolution the basic objective of MD simulations stayed the same, namely to propagate a given molecular system through time to obtain a fine grained trajectory of the dynamics. The length of today's trajectories is only obtainable by treating the molecular system classically (as an ensemble of balls connected by springs) and disregarding all quantum-mechanical effects. This in turn leads to the necessity of parametrisation of otherwise non-tractable interactions, these parameters are commonly bundled into a so called force-field which, in the case of the AMBER gaff force field, has the form[135]

$$U = \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} \frac{\upsilon_n}{2}[1+\cos(n\phi-\gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{\mathbf{r}_{ij}^{12}} - \frac{B_{ij}}{\mathbf{r}_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}}\right] \quad (40)$$

where K_b, K_θ and υ_n are the force constants for the bond, angle and dihedral terms respectively; b_0, θ_0 and γ are equilibrium bond, angle and dihedral parameters; and *n* is the multiplicity. For the nonbonded part of the potential the *A*, *B*, and *q* parameters have to be determined and r_{ij} is the distance between particle *i* and *j*. Parametrisation and refinement of these force field parameters is still a branch of active research, with newer trends being the inclusion of polarisable terms into the electrostatic interactions, allowing bond breaking to occur and so on.[136–138] After a force field is established the next step is to calculate the forces that are acting on the system as

$$\mathbf{F}_i = -\nabla U(\mathbf{r}_i), \quad (41)$$

where here the force on particle *i* is defined as the negative gradient ($-\nabla$) of the potential. When the forces are known Newton's second law of motion,

$$\mathbf{F}_i = m_i \mathbf{a}_i, \quad (42)$$

can be used to link the potential to the dynamic property of acceleration, where the force *F* acting on a particle *i* being written as the product of the mass *m* and acceleration *a*. To propagate the system in time an integrator is used, with the most common one being the Verlet algorithm.[139] The Verlet algorithm is derived from a Taylor series expansion to the second order term around the current positions of the particles in positive direction ($+\Delta t$):

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \frac{\mathrm{d}\mathbf{r}(t)}{\mathrm{d}t}\Delta t + \frac{\mathrm{d}^2\mathbf{r}(t)}{\mathrm{d}t^2}\frac{\Delta t^2}{2} + \ldots \quad (43)$$

and negative direction ($-\Delta t$):

$$\mathbf{r}(t-\Delta t) = \mathbf{r}(t) - \frac{\mathrm{d}\mathbf{r}(t)}{\mathrm{d}t}\Delta t + \frac{\mathrm{d}^2\mathbf{r}(t)}{\mathrm{d}t^2}\frac{\Delta t^2}{2} - \ldots \quad . \quad (44)$$

Combining and restructuring equations (43) and (44) results in the final Verlet algorithm,

$$\mathbf{r}_i(t+\Delta t) = 2\mathbf{r}_i(t) + \mathbf{a}_i(t)\Delta t^2 - \mathbf{r}_i(t-\Delta t), \quad (45)$$

where the velocities of equation (43) and (44) cancel out. Modern molecular dynamics codes mostly implement the Verlet algorithm in one of the algebraically identical forms known as velocity Verlet[140] or leap frog.[141]

The relative straightforwardness of MD simulations is a double-edged sword because it gives people a false sense of safety in regard to the usage of this technique. That is why particular consideration regarding the used water model is warranted, but too often neglected. The most commonly used water models are TIP3P[142] and SPC/E,[143] which are described by three interaction sites and were parametrised to reproduce specific physical observables. The "performance" especially of the TIP3P model is one of the worst,[144] in contrast to one of the best-performing water models to date, the TIP4P/2005[144] water model, which has a fourth off-centre point charge. The general recommendation would be to use the TIP4P/2005[145, 146] water model: This can unfortunately lead to inconsistencies, because almost all modern protein force fields were parametrised with either TIP3P or SPC/E. Therefore it is often advisable to use these instead, to circumvent inconsistencies within the simulation, or for that matter the 3D RISM calculation. For 3D RISM calculations, another problem is that a four site water model is intractable, which is likely due to numerical issues.[147]

2.4.1 Free energy calculations and error estimation

There are plenty of free energy estimation methods available ranging from scoring algorithms to fully atomistic molecular dynamics simulations, where the user has to do the trade-off between speed (scoring algorithms based on empirical functions) and accuracy (fully atomistic MD simulations). The main property of free energies, which is exploited in MD simulations, is that they are state functions, and thus their calculation or measurement is path independent and can be done through artificial, or alchemical routes. One of the widely used methods to estimate relative association free energies with MD simulations is thermodynamic integration (TI), derived by Kirkwood in 1935.[148] The first step is to link the potential energy of the system to a coupling parameter λ in the form

$$U(\mathbf{r},\lambda) = (1-\lambda)U_{\mathrm{A}}(\mathbf{r}) + \lambda U_{\mathrm{B}}(\mathbf{r}), \tag{46}$$

where A corresponds to the end state of the system and B to the starting state of the system. Following this, the Helmholtz free energy A can be written as

$$A(\lambda) = -kT \frac{Z_Q(\lambda)}{V^N}, \tag{47}$$

here $Z_Q(\lambda)$ is the partition function and V^N is the volume of the system to the power of N, the particle number. The derivative of equation (47) with respect to λ yields, after some simple algebra:

$$\begin{aligned} \frac{\partial A(\lambda)}{\partial \lambda} &= -kT \frac{1}{Z_Q(\lambda)} \frac{\partial Z_Q(\lambda)}{\partial \lambda} \\ &= -kT \frac{1}{Z_Q(\lambda)} \int \exp(-\beta U(\mathbf{r},\lambda)) \left(-\frac{1}{kT}\right) \frac{\partial U(\mathbf{r},\lambda)}{\partial \lambda} \mathrm{d}\mathbf{r}. \\ &= \left\langle \frac{\partial U(\mathbf{r},\lambda)}{\partial \lambda} \right\rangle_\lambda \end{aligned} \tag{48}$$

Through integration of the ensemble average over λ, the Helmholtz free energy can be computed or numerically approximated by the sum over incremental λ values:

$$\Delta A = \int_0^1 \left\langle \frac{\partial U(\mathbf{r},\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \approx \sum_i \left\langle \frac{\partial U(\mathbf{r},\lambda)}{\partial \lambda} \right\rangle_{\lambda_i} \Delta\lambda_i . \tag{49}$$

Here it is worth noting that the Helmholtz free energy is equivalent to the Gibbs free energy, if the volume of the system (*NVT*) is equal to the corresponding pressure in the *NpT* ensemble. The conceptually similar and as important technique of free energy perturbation shall be mentioned here for the sake of completeness.[149] For some modern applications and reviews regarding free energy calculations the reader is referred to Ref. [150–152].

A priori it is not possible to make sure that every frame in a MD simulation is statistically independent, but it is possible to estimate the correlation time between MD frames after the simulation has been run by several means. Amongst others blocking analysis is a viable option, which additionally allows to correct the calculated error of the estimated observable. In blocking analysis the statistical inefficiency *s* is written as

$$s = \lim_{\tau_b \to \infty} \frac{\tau_b \sigma^2\left(\langle T \rangle_b\right)}{\sigma^2(T)}, \tag{50}$$

where τ_b is the block length, $\sigma^2\left(\langle T \rangle_b\right)$ is the variance of block T and $\sigma^2(T)$ is the variance of the whole dataset.[153] First the block averages have to be calculated by

$$\langle T \rangle_b = \frac{1}{\tau_b} \sum_{\tau=1}^{\tau_b} T(\tau), \tag{51}$$

where the trajectory is split into blocks of length τ_b and the number of blocks $n_b = \tau_{all}/\tau_b$, with τ_{all} being all time steps in the trajectory.[153] With the computed block averages calculated, the block variance can be estimated through:[153]

$$\sigma^2\left(\langle T\rangle_b\right) = \frac{1}{n_b}\sum_{b=1}^{n_b}\left(\langle T\rangle_b - \langle T\rangle_{all}\right)^2 . \tag{52}$$

For practical purposes it can be helpful to estimate the statistical inefficiency not with classical limit value consideration, but to take the maximum obtained value of *s*.

2.5 Machine learning techniques

2.5.1 Deep feedforward networks

The area of machine learning can be generally divided into supervised and unsupervised learning. Among those two a zoo of methods exists which have advantages and disadvantages. In unsupervised learning, the chosen method has to be able to extract a function from unlabelled data, which means that for a given distribution the *y*-value is unknown but has to be inferred from the data itself. An example for that would be principal component analysis.[98, 154] Supervised learning on the other hand works with labelled data, which means that it works on pairs of *x*- and *y*-values, and tries to find a function which allows to map new examples to the correct classes (classification tasks) or numerical values (regression tasks).[98, 155] As of this writing neural networks and in particular deep neural networks are one of the most hyped machine learning techniques,[155, 156] and are already used in the field of chemistry.[157–161]

Although the success of neural networks and deep learning (as it is called today) kicked off in the late 20th until the early 21st century, these techniques have already got a long history.[156] The first models, which were inspired by the structure of the brain and the way learning in a

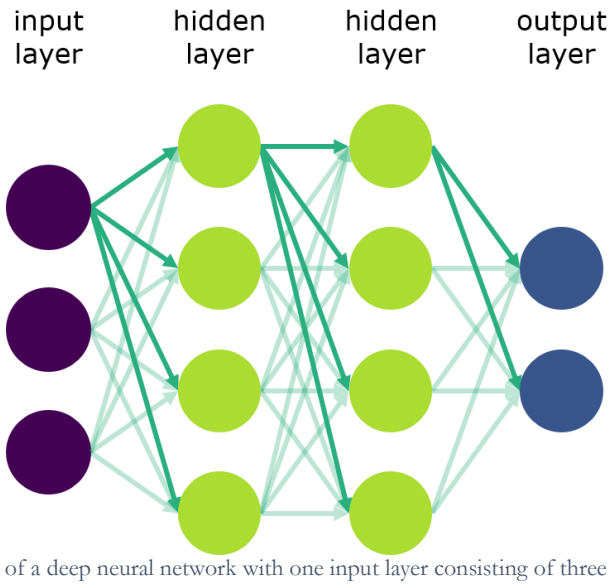

Figure 4: Basic representation of a deep neural network with one input layer consisting of three input neurons (purple circles), two hidden layers, consisting of four hidden neurons each (green circles) and one output layer consisting of two output neurons (blue circles). The arrows represent the weights and the neurons are connected in a dense fashion, which means every neuron is connected to all neurons of the next layer.

biological setting could work, were derived by McCulloch and Pitts in 1943[98, 162, 163] where the weights of the model were not learned but had to be assigned by a human operator. After that, the first trainable single neuron model was devised by Rosenblatt in 1953.[98, 163, 164] These models were then enhanced in the so-called second wave of neural networks around 1980 – 1995 where the training of models with one or two hidden layers became possible by the means of "back propagation" (details see below).[98, 156, 163, 165, 166] At the moment we are right in the middle of the so-called third wave of neural networks, which is now often called "deep learning," and began around 2006 with the work of Hinton *et al.* on "Deep belief networks".[98, 156, 167] The third wave was also sparked and supported on the hardware side by the exploitation of "General-purpose computing on graphics processing units" (GPGPU).[98] This can also be seen in the high rate of GPGPU capabilities, baked into every major deep learning library on the market, e.g. TensorFlow,[168] Theano,[169–171] Caffe[172] and many more.

In a nutshell, a neural network is composed of so-called neurons and the connections between these neurons (see Figure 4) which are arranged in a layer-wise fashion,[98, 155, 156] starting with the input layer that is followed by one or more hidden layers and an output layer. Mathematically speaking, a neural network can be described as a chain of functions: $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$ where $f^{(1)}$ represents the first layer, $f^{(2)}$ the second layer, and so on.[98] The training process aims to drive $f(x)$ to approximate $f^{*}(x)$ (which is the underlying function describing the true distribution of x as good as possible.[98, 155] The flow of information through the network can thus be described through:

$$\mathbf{x} \rightarrow \mathbf{h} \rightarrow \mathbf{y}\,. \tag{53}$$

The feature matrix (or vector depending on the problem at hand) **x** enters the network than the aforementioned transformations described in equation (57) and (58) are applied in the hidden units of the network and at last the predicted **y** values leave the network.

The training data consists of pairs of (possibly noisy) x corresponding to $y^* \approx f^*(x)$.[98] In detail, a hidden neuron takes the incoming input vector **x** and applies the following transformation in the form

$$\mathbf{u} = \mathbf{W}^{\mathrm{T}}\mathbf{x}\,. \tag{54}$$

Here, **u** is the resulting vector $\mathbf{W}^{\mathrm{T}}$ is the "weight" matrix (which is adjusted during the training process). After that, the "activation" function $a(\mathbf{u})$ is applied elementwise to the **u** vector: [100, 159, 167]

$$\mathbf{h} = a(\mathbf{W}^{\mathrm{T}}\mathbf{x})\,. \tag{55}$$

The choice of the activation function a can have a significant impact on the learning process and the achievable performance of the neural network[173]. One of the default choices today is the "rectified linear unit"[174, 175] defined as $\mathrm{a}(\mathbf{u})=\max(0,\mathbf{x})$. Other possible activation functions are the hyperbolic tangent or "sigmoid" function.

The functional form of the output units of a neural network are often depending on the problem at hand. For binary classification tasks a "sigmoid unit" can be a good choice for instance.[98] For multi-class classification tasks, the "softmax unit" is often used. For regression tasks a linear output unit is often a sensible choice. These functions are chosen because they output values between zero and one, which is preferable for classification tasks.

In supervised learning the cost, which is a measure of the difference between the calculated y ($f(x)$) values and the "true" y^* ($f^*(x)$) values, is calculated by the loss function. For classification tasks, one of the possible loss functions (and often a very sensible choice) is the cross entropy, or frequently called the log-loss, which is defined as:[98]

$$L(\mathbf{y}^*, \mathbf{y}) = -\sum_i y_i^* \log y_i\,. \tag{56}$$

For regression tasks typical loss functions include, but are not restricted to, the mean absolute error or the mean squared error.[98] During the training process the cost is minimized to reflect $f^*(x)$ as well as possible. This is often done using an optimization technique called stochastic gradient descent (see equation (61)), or variants thereof. The actual training process of a neural network can be split into three parts: The first step is the forward propagation of the data through the network

$$\mathbf{t}^{(l)} = \mathbf{W}^{(l)}\mathbf{h}^{(l-1)}, \tag{57}$$
$$\mathbf{h}^{(l)} = a(\mathbf{t}^{(l)}). \tag{58}$$

Here l is an index that runs from the first layer to the last layer. The matrix **h** stores the values for every node, **h** equals **x** in the input layer and **y** in the output layer.[98] Then the loss $L(\mathbf{y}^*, \mathbf{y})$ has to be computed. In the second step the gradient of the loss has to be computed with regard to the weights. In the rather simple case, as described here, where the weights are the only parameters that can be varied, makes it clear that a neural network can be interpreted as a chain of functions. Therefore, the calculation of the gradient relies heavily on the chain rule of calculus. The algorithm of choice is called back propagation[98, 163, 165, 166] and can be written as

$$\mathbf{g}^{(\text{out})} = \nabla_{\mathbf{y}} L(\mathbf{y}^*, \mathbf{y}), \tag{59}$$
$$\mathbf{g}^{(l)} = \nabla_{\mathbf{h}^{l-1}} L = \mathbf{W}^{(l)\mathrm{T}}\mathbf{g}^{l-1} \circ a'(\mathbf{h}^{l}). \tag{60}$$

First, the gradient of the loss in the output layer has to be computed, this is shown in equation (59), which is done by calculating the Jacobian with regard to the output values **y**. In the next step the gradient for every layer can be derived by following equation (60). Here $\mathbf{g}^{(l)}$ is the gradient in layer l, $\mathbf{g}^{(l-1)}$ is the gradient in layer l -1, $a'(\mathbf{h}^{l})$ is the derivative of the activation function with regard to the weights and "$\circ$" is the Hadamard product. The last training step now consists of the update of the weight matrices **W** and in the case of stochastic gradient descent, it can be written as:

$$\mathbf{W}^{l,\text{new}} = \mathbf{W}^{l,\text{old}} - \lambda\mathbf{g}^{(l)} \tag{61}$$

where λ is the "learning rate" which can be fixed or variable. The term stochastic in this case can be attributed to the fact that the gradient is not computed over the whole dataset, but randomly drawn small "mini" batches. Therefore stochastic gradient descent is often described

as (mini-) batch gradient descent. Often the Adam[176] optimizer is used, which is a variant of the gradient descent method.

2.5.2 Gradient boosted trees

Gradient boosted trees and variants thereof are one of the top performing machine learning techniques.[99] They are also successfully used in molecular modelling.[37, 177] In a comparison Ashtawy[177] *et al.* did among machine-learning-based and classical scoring functions boosted regression trees were also one of the top performers.[37, 177] In 2016, Chen *et al.* published a paper (initial release of the software was in 2014) about their boosting algorithm, named XGBoost (abbr. for extreme gradient boosting) which is based on the gradient boosting model of Friedman.[178] XGBoost is widely recognized in the data science community which is shown by the adoption in many challenges and cups. For example, in the 2015th KDDCup, every winning team in the top 10 used XGBoost.[99] Furthermore, on the data science competition website "kaggle" 17 of the 29 winning solutions during the 2015 timeframe used XGBoost.[99] Of these 17 solutions, 8 solely used XGBoost with the rest using a combination of XGBoost and neural networks.[99] As far as the author knows XGBoost was never used in a molecular modelling context specifically.

XGBoost is a combination of several well proven techniques in machine learning and to approach it in a more accessible manner the key components are first described alone and at the end of the chapter the actual XGBoost method is introduced.

At first "boosting" as a technique, which can be seen independently of the used regression or classification method, is introduced and can be described in the following manner.[158, 167, 184, 185] The boosting algorithm uses an ensemble of independent regression models and starts by assigning weights to each individual training sample x_i:

$$x_{i,w} = w_i x_i \,. \tag{62}$$

Here, the weights w_i are initialised with $1/N$, where N is the number of samples and $i = 1,2,\ldots,N$. Then a model is fitted (which is described below) using the weighted training examples $x_{i,w}$ which yields an estimated y value.[154, 163, 179, 180] This is followed by the computation of the error rate for the regression model k,

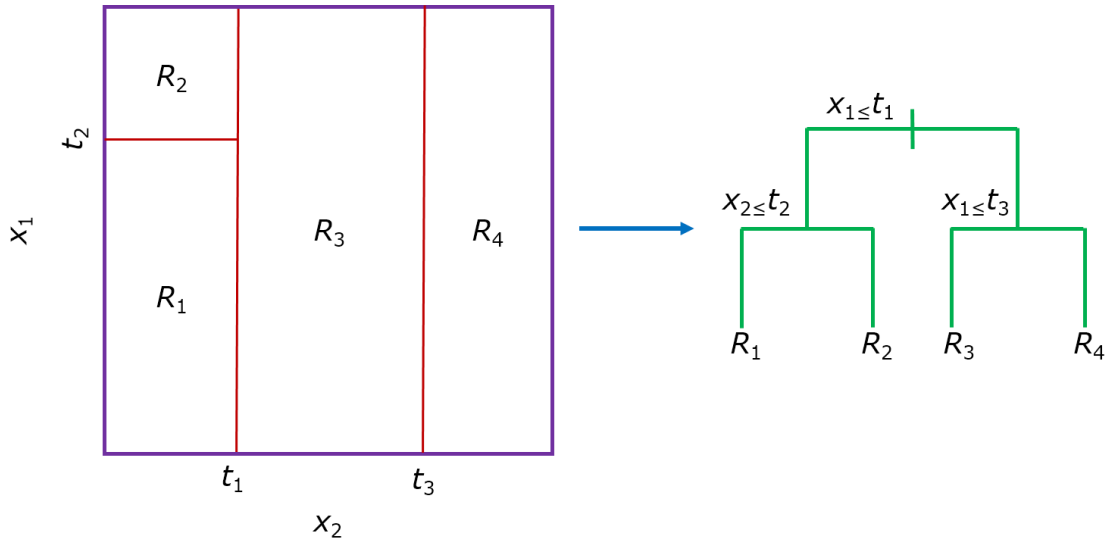

Figure 5: Left side shows the partition of a two-dimensional feature space obtained through recursive binary splitting. The right side shows the corresponding tree structure. Graphic adapted from "The Elements of Statistical Learning"[180]

$$\mathrm{err}_k = \frac{\sum_{i=1}^{N} w_i L\left(y_i^*, y_i\right)}{\sum_{i=1}^{N} w_i}, \tag{63}$$

where $L\left(y_i^*, y_i\right)$ is the loss (a measure for the difference) between the prediction and the true y_i^* value, the error rate thus scales with the loss.[154, 163, 179, 180] This is followed by the computation of the parameter α_k, which is defined as:

$$\alpha_k = \log\left(\left(1-\mathrm{err}_k\right)/\mathrm{err}_k\right). \tag{64}$$

Then the weights w_i are adjusted with

$$w_i = w_i \exp\left[\alpha_k L\left(y_i^*, y_i\right)\right], \tag{65}$$

which then allows to calculate the final "boosted" prediction as:

$$f(x) = \sum_{k=1}^{K} \alpha_k f_k(x). \tag{66}$$

Here $f_m(x)$ is the prediction of the individual regression models that are part of the ensemble and M is the number of regression models used.[154, 163, 179, 180] The benefit of the "boosting" technique lies in the fact that the weights for samples that are misclassified or have a high loss are growing exponentially. The consequence is that after every round of "boosting" the algorithm pays more and more "attention" to samples that are misclassified or have a high loss.

Tree models can be described graphically as seen in Figure 5. Under the premise that the problem at hand is a regression problem with continuous y values and the two features (data that describes the y values) x_1 and x_2 one can split the feature space by t_1, t_2, and t_3, which yields the respective regions R_1, R_2, R_3, and R_4.[180] Then the function $f(x)$ (the predicted y) would be:

$$f(x) = \sum_{m=1}^{4} c_m I\{(x_1, x_2) \in R_m\}, \tag{67}$$

with the function I being one if the pair (x_1, x_2) belongs to region m and zero in all other cases.[180] It can be shown that the best c_m is given by

$$c_m = \text{ave}\left(y_i \mid x_i \in R_m\right), \tag{68}$$

which is the average of y_i in region R_m.[180]

The training algorithm now has to decide how big the "trees" are allowed to get, which it controls through "pruning" (finding a minimal effective tree), and has to optimize the tree structure in a way that represents the function $f(x)$ in an optimal way. Regression trees alone are often called "weak" or "base" learners.

The XGBoost algorithm which combines gradient boosting with regression trees tries to combine an ensemble of "weak" or "base" learners (e.g. separate decision trees) into a "strong" learner, which means a model with good predictive capabilities and can be described in the following way.[56, 163, 179, 181] A tree ensemble model can be written as

$$y_i = \sum_{k=1}^{K} f_k(x_i), \tag{69}$$

where y_i is the predicted value and f_k is the number of K additive functions ("trees") ("weak" or "base" learner) that are used to predict the output.[55, 99] It should be noted that every

$$f_k(x) = w_{q(x)}, \tag{70}$$

resembles an independent tree structure q with leaf weights w.[99] For the learning process the following regularized objective can be defined:[99]

$$L = \sum_i l(y_i^*, y_i) + \sum_k \Omega(f_k) \tag{71}$$

where l is a convex loss function and the second term Ω is a regularization term, which is typically the weighted L2-norm, that penalizes the complexity of the model.[99] For the optimization process equation (71) is rewritten into

$$L^{(t)} = \sum_{i}^{n} l(y_i, y_i^{*(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{72}$$

here the index t represents the current iteration and i the current instance.[99] Equation (72) is the optimization target and is trained "greedily" by adding the function f_t that improves the model best.[99]

3 Designing molecular complexes using free-energy derivatives from RISM-*uu*[2]

3.1 Introduction

While RISM-*uu* theory has been used for PMF calculations in the past,[68, 94] its related derivatives have not been tested for design purposes thus far to the best of the authors knowledge. Therefore, first several fundamental questions concerning the accuracy in comparison with related approaches in this proof-of-principle study have to be addressed. The focus therefore lies on the well-studied 18-crown-6 ether complexed with the alkali ions sodium, potassium and caesium in water with various potential parameter sets for the ions. Before computing FEDs based on the *uu* PMF, the quality in comparison with explicit free energy MD simulations using thermodynamic integration (TI) for given fixed relative complex geometries defining a reaction path has to be addressed, which has not been done before. The *uu* and MD data are compared with the PMF computed from *uv* calculations with explicitly placed ions along the pathway in order to draw conclusions about the influence of approximations on the PMF topography. Finally, *uu*-based FEDs for varying ion size parameters are computed and demonstrated to yield consistent and physically reasonable results compared to literature data.

[2] Reused in part with permissions from F. Mrugalla, S. M. Kast, *J. Phys.: Cond. Matter* **2016**, *28*, 344004. © 2016 IOP Publishing Ltd.

3.2 Computational details

Throughout all calculations the identical, rigid 18-crown-6 ether structure was used in its *D3d* symmetry as obtained by geometry optimization with `Gaussian 03` (Rev. D.02) in the gas phase with B3LYP/6-31G*.[182] The interaction parameters of the crown ether were selected from the "optimized potential for liquid simulation" (OPLS) force field (see Table 1)[183, 184] employing Lorentz-Berthelot mixing rules throughout, similar to our earlier 1D RISM work on 18-crown-6 in nonaqueous solvents.[185] For alkali ions, four different K^+ parameter sets for FED calculations were tested, all summarized in Table 1.[186–189] TI MD reference simulations and explicit 3D RISM-*uv* calculations were performed also with other ions, using exclusively the MacKerell *et al.* set.[186] Water was described by the TIP3P model, using for reference MD calculations, the original form[142] and a variant with nonzero Lennard-Jones parameters on hydrogen (σ = 0.4 Å, ε = 0.0459 kcal mol^{-1}). This modification is necessary to avoid singularities implied with 1D RISM iterations and was used for all integral equation calculations including those with susceptibilities taken from MD.

Table 1: Force field parameters of 18-crown-6 and ions.

Atom	q / e	σ / Å	ε / kcal mol^{-1}
C(18-crown-6)	0.14	3.5000	0.0660
O(18-crown-6)	-0.40	2.9000	0.1400
H(18-crown-6)	0.03	2.5000	0.0300
Na^{+} [a]	1.00	2.4299	0.0469
K^{+} [a]	1.00	3.1426	0.0870
K^{+} [b]	1.00	3.0380	0.1937
K^{+} [c]	1.00	4.7360	0.0003
K^{+} [d]	1.00	3.5600	0.1304
Cs^{+} [a]	1.00	3.7418	0.1900

[a]MacKerell *et al.*,[186] [b]Joung and Cheatham,[187] [c]Åqvist,[188] [d]Wipff *et al.*.[189]

1D RISM calculations with the dielectrically consistent (DRISM/HNC) theory[190, 191] for pure water (modified TIP3P) were performed on a logarithmic grid with 512 grid points ranging from 5.98·10^{-3} Å to 164.02 Å. The solvent density was set to 0.0333 Å^{-3}, the temperature to 298.15 K, and the dielectric constant to 78.4. For the MD extraction of the susceptibility function with the same water model, a simulation of a water box with 4033 molecules was set up using `tleap`[192] and equilibrated over 10 ns *NpT* simulation (pressure was 1 bar controlled by the Langevin piston, temperature was 298.15 K via the Langevin method using default

settings) with a 2 fs time step in NAMD.[193] Short range potentials were truncated at 12.0 Å and the particle mesh Ewald (PME) method was employed for treating Coulomb interactions. A frame with minimal deviation from the target density was selected and a *NVT* production run over 20 ns was performed with identical simulation parameters, also for original TIP3P. Pair distribution functions for susceptibility extraction were determined with the Gromacs tools applying a histogram bin size of 0.02 Å on the basis of 20000 frames[194–196] and smoothed up to a maximum distance of 23.88 Å, beyond which susceptibilities were extrapolated by DRISM/HNC, following closely the procedures employed earlier.[197–199] Convergence criteria for 1D RISM calculations were a maximum residual norm of the direct correlation functions of 10^{-7} and 0.00023 for HNC and MD extraction, respectively.

3D RISM-*uv*/PSE-(n = 1-3) calculations at 298.15 K were performed on cubic grids of 200^3 points with a spacing of 0.2 Å. The convergence criterion for the 3D RISM-*uv* calculations was set to 10^{-4} for the maximum residual norm of direct correlation functions. For the PMF calculations one of the ion species (K^+, Na^+, Cs^+) was placed along a 1D path defining the z-axis of the 18-crown-6 (see Figure 6 together with an illustration of direct interaction energies). The path was symmetrically constructed with a sampling rate of 0.2 Å and a maximum distance to the center of the crown ether of 10.0 Å, implying 101 points. In the 3D RISM-*uu* case the 3D RISM-*uv* calculations of the crown ether were reused while for the ions 1D RISM-*uv* calculations were performed using modified TIP3P susceptibilities (convergence threshold 10^{-5}) and interpolated to the 3D grid by cubic splines. FEDs from *uu* calculations were obtained by numerical differentiation with a 5-point stencil ($\Delta\sigma$ = 0.02 Å).

The simulation system for the TI calculations was a cubic box with the rigid 18-crown-6 ether, a bound ion at the centre and 4036 water molecules. First a *NpT* simulation over 1 ns with a 2 fs time step at 0.5 bar was performed, followed by a *NpT* run at 1 bar over 10 ns. TI simulations in the *NVT* ensemble were initiated from the frame with minimal deviation from the target density. For the same ion positions as in integral equation calculations the coupling parameter λ was scaled in $\Delta\lambda$ = 0.1 steps, decoupling ion-solvent and ion-host interactions. For each value of λ the system was simulated for 150000 steps of which 25000 were discarded for equilibration. Appropriate numbers of statistically uncorrelated frames were determined by blocking analysis[153] before calculating the PMF by numerical integration of cubic spline interpolants. A similar protocol was used for decoupling the ion in the absence of the crown ether for defining the reference state, whereby artifacts attributed to the presence of a net charge

in TI simulations effectively cancel. After the TI calculations of all 101 points the symmetric setup was exploited and the average PMF of the corresponding points from both sides of the path computed. For K^+ the relevance of incomplete sampling (hysteresis effect) by repeating the TI simulations in the reverse direction was checked, *i.e.* by recoupling interactions starting from the final decoupled states.

The crown ether picture in Figure 6 has been generated using the `PyMOL` software.[200] All other plots have been created using the software `Gnuplot`.[201] Integral equation calculations have been performed with software developed in our laboratory. Data analysis was done using `Mathematica`.[202]

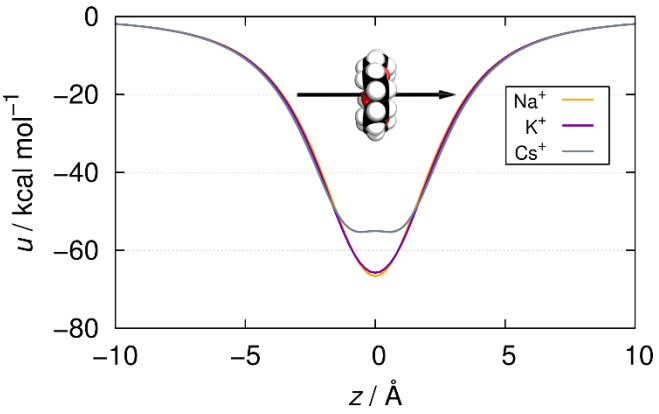

Figure 6: Schematic representation of 18-crown-6 and the chosen ion translocation path, with direct interaction energies between crown ether and the various ion species using the MacKerell et al. model[186] along the *z* axis. In this case, the space of relative distances and orientations ($\mathbf{R}_{12}$, $\boldsymbol{\Omega}_{12}$) is reduced to three dimensions and one component (z) of the distance vector has been singled out.

3.3 Results and discussion

First the accuracy of both integral equation approaches to the PMF has to examined, explicit super-molecule calculations by 3D RISM-*uv* and the most efficient 3D RISM-*uu* estimate, in comparison with TI reference results, which is shown in Figure 7 for 18-crown-6 with K^+ using the MacKerell *et al.* parameters.[186] In the top left panel, only *uv* and *uu* results for various PSE orders (also applied to the underlying *uv* calculations in the *uu* case) on the basis of DRISM/HNC(*vv*) water susceptibilities are depicted. While the overall topography of the free

energy surfaces along the chosen path appears to be similar, the absolute heights of barriers and the depth of minima are considerably different. Notably, the location of the global minimum is identical for all methods whereas the precise locations of the barriers differ slightly. More problematic for quantitative applications is the lack of significant free energy barriers in the *uu* case in general, while apparently explicit *uv* and *uu* results tend to converge toward better agreement with increasing PSE order. For the barriers the difference between the explicit *uv* and *uu* data is the largest for the PSE-1 ($\Delta w \approx 6.8$ kcal mol^{-1}) and the smallest for PSE-3 ($\Delta w \approx 4.1$ kcal mol^{-1}). For the minima the differences between the two calculation methods are generally smaller, and, similar to the barriers, they decrease with increasing PSE order ($\Delta w \approx 3.7$ kcal mol^{-1} for PSE-1 to $\Delta w \approx 0.9$ kcal mol^{-1} for PSE-3. Yet, the PSE order has more significant impact on minima than on barriers.

The bottom left panel now shows *uv* (PSE-2) and *uu* data also with the MD-extracted water susceptibility, and in comparison with reference (forward) TI simulations. Using MD-generated

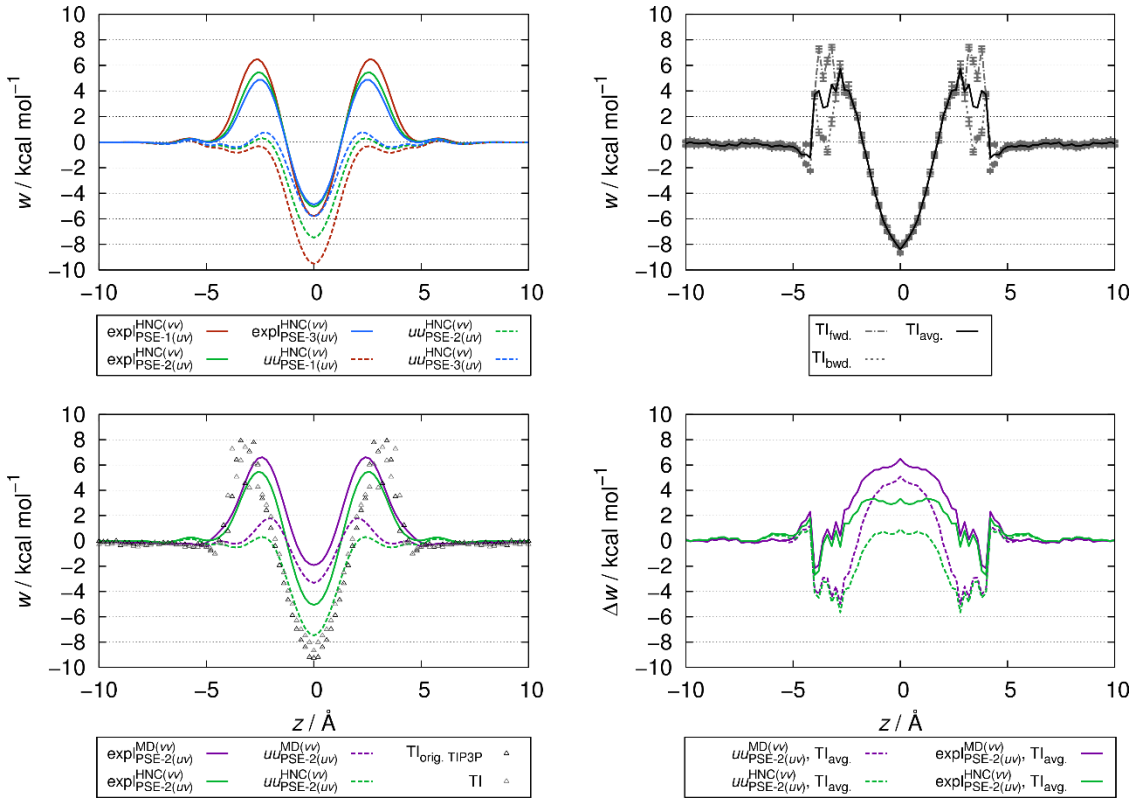

Figure 7: Comparison of the K^{+}[186] PMFs with 18-crown-6 in water for various calculation setups. Top left: explicitly ("expl") placed ion within 3D RISM-*uv* calculations and *uu* results with underlying *uv* PSE order given as subscript, DRISM/HNC(*vv*) pure solvent susceptibility. Bottom left: *uv* and *uu* results with DRISM/HNC(*vv*) and MD-extracted ("MD(*vv*)") pure solvent susceptibility in comparison with reference TI simulation data employing the original and modified TIP3P water models. Top right: forward and backward TI data with statistical error bars from blocking analysis, averaged over left and right half, along with average TI PMF (modified TIP3P model), indicating a hysteresis effect. Bottom right: deviation of PMFs from various *uv* and *uu* approaches from average TI data.

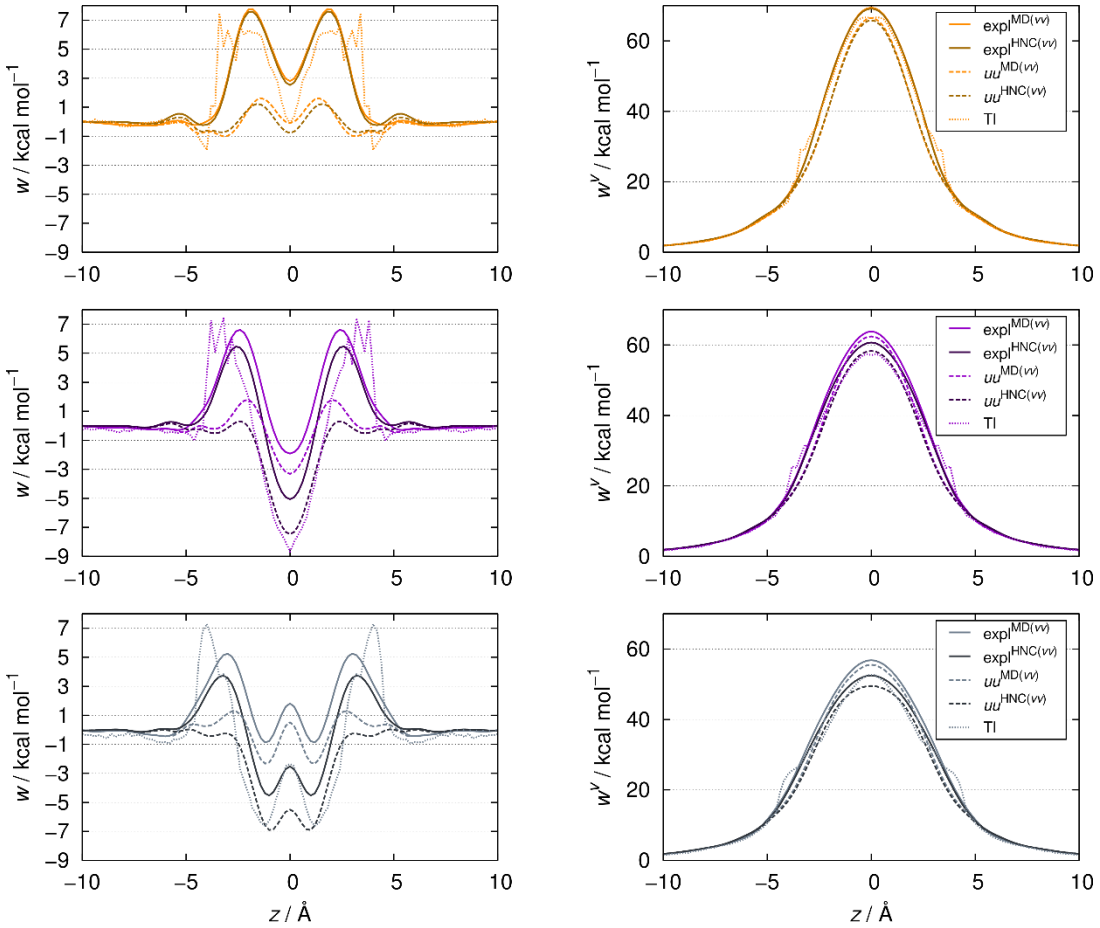

Figure 8: PMF (left column) and desolvation penalty (right column) for Na^+ (yellow, top), K^+ (purple, middle), and Cs^+ (grey, bottom); explicit *uv* and underlying *uv* calculations for *uu* with PSE-2 and DRISM/HNC or MD-extracted susceptibilities, all including TI reference with the modified TIP3P water model and the MacKerell et al. parameter set[186].

χ functions generally has opposing effect on barriers and on minima. While barriers benefit somewhat from using MD input in comparison with TI, the effect on minima is less pronounced, yet visibly tending in the more strongly deviating direction. However, in absolute numbers the agreement between best *uu* and *uv* setups with reference TI data is reasonable, Δw^{uu} < 1 kcal mol^{-1} and $\Delta w^{uv} \approx 4$ kcal mol^{-1} for HNC(*vv*) in comparison with TI using the modified TIP3P model as in integral equation calculations. For barriers, the discrepancies are $\Delta w^{uu} \approx 5$ kcal mol^{-1} and $\Delta w^{uv} \approx 1$ kcal mol^{-1} for MD(*vv*). Note that the precise location of the barriers in the TI case is shifted to slightly larger distances to the crown center, and that the difference between original and modified TIP3P models is negligible such that only TI data with the latter model was chosen as reference for following analyses.

The top right panel of Figure 7 reveals substantial TI artifacts arising from strong hysteresis effect upon repeating the simulation in the reverse coupling parameter direction. These amount to around 2-3 kcal mol^{-1} around the barrier regions, much larger than the statistical sampling

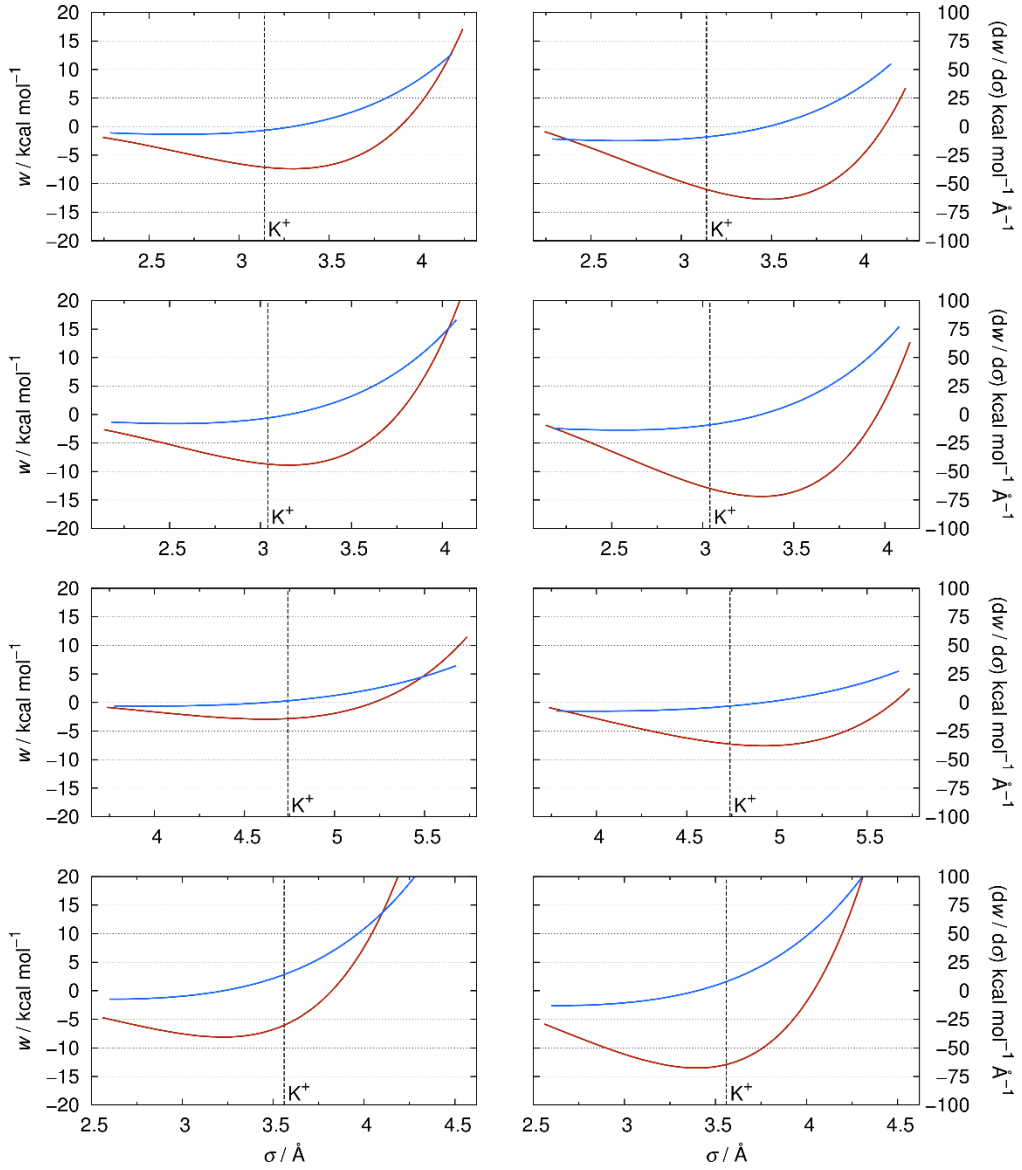

Figure 9: FEDs (blue, right ordinate) and PMF curves (red, left ordinate) for various parameter sets for K^+, placed at the crown center, on the basis of uu calculations (PSE-2 basis for underlying *uv* data). The original K^+ value for the respective parameter set is indicated by the dashed line. From top to bottom: MacKerell et al.[186], Joung and Cheatham[187], Åqvist[188], Wipff et al.[189] parameter sets; left/right columns show data with MD-extracted and DRISM/HNC susceptibilities, respectively.

error. The origin is related to the special simulation setup that was employed, namely the choice of fixed ion positions relative to the crown ether. Visual inspection of trajectories shows for ion distances slightly above or below the ring center that water molecules cannot sufficiently sample the narrow regions between host and guest. The true TI barrier height is likely closer to the *uv* data than estimated from the forward direction only. This also becomes clear from the differences between explicit *uv* or *uu* PMFs and the hysteresis-averaged TI curve shown in the bottom right panel of Figure 7, which corresponds to the difference of desolvation penalties (see also below for a more detailed description) since the host-guest interaction energies are

identical for all approaches. This demonstrates more clearly that *uu* deviates topographically similarly but systematically underestimates the desolvation work compared to *uv*, while *uv* mostly overestimates the penalty compared to TI. Near the center, *i.e.* in the region of largest error, the integral equation artifacts have therefore fortuitously less negative impact on absolute numbers. This coincidental cancellation of errors near PMF minima is, however, a general pattern found for all ions examined, as shown below in Figure 8.

A comparison of 3D RISM-*uu* with *uv* and TI calculations for Na^+, K^+, Cs^+ sheds more light on the origin of the apparent discrepancies. Besides the PMF the desolvation also show penalties for the three ions which are simply the differences between PMF and underlying direct interaction energies. This penalty is a measure for the free energy impact of stripping water molecules from the ions when entering the crown ether. In general the desolvation penalty calculated by *uu* and *uv* yields a mixed picture in regard to the TI reference, with *uu* calculations based on DRISM/HNC(*vv*)[190, 191] performing systematically best near the PMF minima. All 3D RISM flavors are capable of reproducing the overall topography of the TI PMF curves. Explicit *uv* calculations tend to be better at reproducing the shape and height of the barriers, while we find, consistently for all ions, *uu* from DRISM/HNC(*vv*)[190, 191] is better suited for the prediction of the depth of the minima. The most efficient approach is therefore an interesting candidate for replacing MD-based binding free energy predictions by an integral equation model. While the PMF topographies for Na^+ and K^+ are similar, the situation differs strongly for Cs^+. Indicated by a local maximum at the center, Cs^+ does not fit into the crown ether cavity and mostly sits on top of it. This result agrees with quantum-chemical calculations and experimental data for this system[203–205] and is nicely reproduced by all RISM-based methods as found from TI.

The analysis presented here indicates the reason why liquid state theories can have difficulties with respect to quantitative predictions. Overall, the agreement between integral equation and TI penalty curves is good, while the precise locations of the onset of the desolvation process differ only slightly. However, the large slope of the desolvation penalty is compensated by a similarly large, opposite slope in the direct interaction energy in this region. This means that two steep, opposing trends can have very large impact on the absolute numbers when added, giving rise to stronger discrepancies in total PMFs as would be expected from separate components. Any improved liquid state theory therefore has to account for an improved description of ion-water distribution functions at close contact, which represents a

considerable challenge for future developments. Such attempts are certainly worthwhile since the computational demand varies widely, by several orders of magnitude, among the methods presented here. For a single state-of-the art processor core and a given relative configuration roughly 16.000 min for TI/MD is needed, 100 min for explicit *uv* calculations, while a *uu* calculation with precomputed *uv* data for separate partners requires only 0.01 min

Turning finally to the FEDs from *uu* calculations, various ion parameter sets for K^+ were tested with respect to the robustness and plausibility of the predictions. Since the PMF topographies between TI and integral equation results differ only slightly and, in particular, the location and depth of the minimum is well described by *uu* theory, good correspondence is expected with results obtained by others who required much higher computational cost. Figure 9 shows FED results with respect to the ion size parameter for various setups. Notably, absolute numbers are strongly influenced by switching between DRISM/HNC and MD-extracted water susceptibilities but not to the same extent the location of the zero-crossing of the FED or the minimum of the PMF, respectively (with the exception of Åqvist set,[188] see further discussion below). In this sense, *uu*-based FEDs represent indeed a robust quantity for optimizing chemical parameters by providing direction information to the molecular designer.

The two bottom panels show positive free energy derivatives for the calculations with the Åqvist[188] (for MD-extracted susceptibilities only) and Wipff *et al.*[189] parameter sets at the original σ value of the potassium ion. These results are in agreement with the study of Cieplak *et al.*[80] where a molecular dynamics study in conjunction with free energy derivatives and the same parameter set (Åqvist[188]) yielded also positive free energy derivatives for the respective σ value. This indicates that the optimal binding partner of the 18-crown-6 ether is an ion with a slightly smaller radius than the original K^+ parameter. At first sight, the data for the other two parameter sets (top panels, MacKerell *et al.*,[186] Joung and Cheatham[187]) seem to contradict this conclusion since they suggest increasing the size parameter. However, for absolute numbers the trends agree with the Wipff *et al.*[189] tendency to yield an optimal σ parameter of around 3.4 Å. Only the Åqvist[188] set appears to deviate in terms of absolute numbers, which is not unexpected since the absolute values for this set are mere fit parameters to represent solvation free energies reasonably, sacrificing any physical meaning. For the other three sets, the overall prediction of an optimal K^+ size appears to be robust and practically independent of the accompanying well depth parameter defining the Lennard-Jones potential.

3.4 Concluding remarks

In this proof-of-principle study it was shown that it is indeed straightforward and physically reasonable to employ 3D RISM-*uu* theory for the purpose of predicting design directions for certain interaction parameters defining variations in chemistry. This investigation was footed on a thorough comparison of the relative strengths and deficits of various integral equation formulations and their inherent dependence on input parameters such as pure solvent data and closure approximations. The benchmark data for this purpose was provided by explicit free energy molecular dynamics simulations based on the same interaction potential and structural model as used in integral equation calculations. Such an analysis rigorously revealed the deficits of a 3D RISM treatment with currently available approximations. In particular, the subtle interplay of opposing quantities, interaction energy and (de)solvation contribution to the total PMF, strongly depends on the physical level of accuracy that defines a liquid state theory. Therefore, much work has to be done to improve those theories to reach quantitative agreement with explicit simulations consistently over PMF landscapes. However, the results also showed that PMF topographies, which are most relevant properties for employing free energy derivatives in practical design work, are reasonably robust and less influenced by the inherent approximations. Hence, even a computationally very efficient theory such as 3D RISM-*uu*, that does not account for higher-order correlations between two solute partners and the solvent as compared to 3D RISM-*uv* on super-molecules, can be envisioned to be developed into a practically useful design model for more complex problems such as protein-ligand binding, which is the topic of the next chapter of this work.

Building on these results the next chapter will apply FEDs calculated with 3D RISM-*uu* to two protein-ligand system. The model will also be extended and verified further.

4 Free energy derivative guided-drug design with RISM-*uu*

4.1 Introduction

In order to apply free energy derivatives to protein ligand complexes in a meaningful manner it is desirable that per atom information can be obtained to drive the decision process. The general idea would be to calculate the free energy derivative of every ligand atom at its respective position in three dimensional space and map this information back to the ligand structure to obtain a picture as in Figure 21. Here, the following central questions arise: (1) is placing the ligand atoms at their original three dimensional position into the *apo* structure of the protein binding site enough (calculations done in this way are denoted by "*apo*" as superscript)? (2) Should the ligand atoms be placed in a one-to-one manner so that the binding site is partially filled with the remaining part of the ligand (calculations done in this way are denoted by "part *holo*" as superscript)? (3) Which 3D RISM parameters should be used to obtain fast and reliable results? All of these question are addressed in this chapter. (4) Are the computed FEDs and derived quantities in accordance to the experimental results, despite all the approximations that are made?

The molecular systems under scrutiny are the kinase domains of "rearranged during transfection" (RET) in complex with AD80 and tRNA guanine transglycosylase (TGT) in complex with an aminoquinazolin derivate.

4.2 Computational details

4.2.1 Structure preparation

All used structures were equilibrated with MD simulation, if not explicitly stated otherwise. In the case of "rearranged during transfection" RET, a homology model was generated by Justina Stark. The modelling and equilibration process being described in detail in Ref. [206] in the following only a brief summary is given: The RET DFG-out model was generated with `Modeller`[207] using VEGFR (pdb: 2OH4) as template. After that, the system was parametrized using the ff99SB force field from `Amber12`[192] for the protein and `GAFF` 1.5[212, 213] for the ligand called AD80 (see Figure 10 A, C). Partial charges for the ligand were calculated using the AM1-BCC method.[210, 211] Then solvated in TIP3P[142] water and neutralized with chloride ions. This process was followed by three successive fully atomistic MD simulations with `NAMD`:[193] starting with a 4 ns long simulation with restraints (only C^{α} atoms) in the *NVT* ensemble followed by a 4 ns long simulation with restraints (only C^{α} atoms) in the *NpT* ensemble

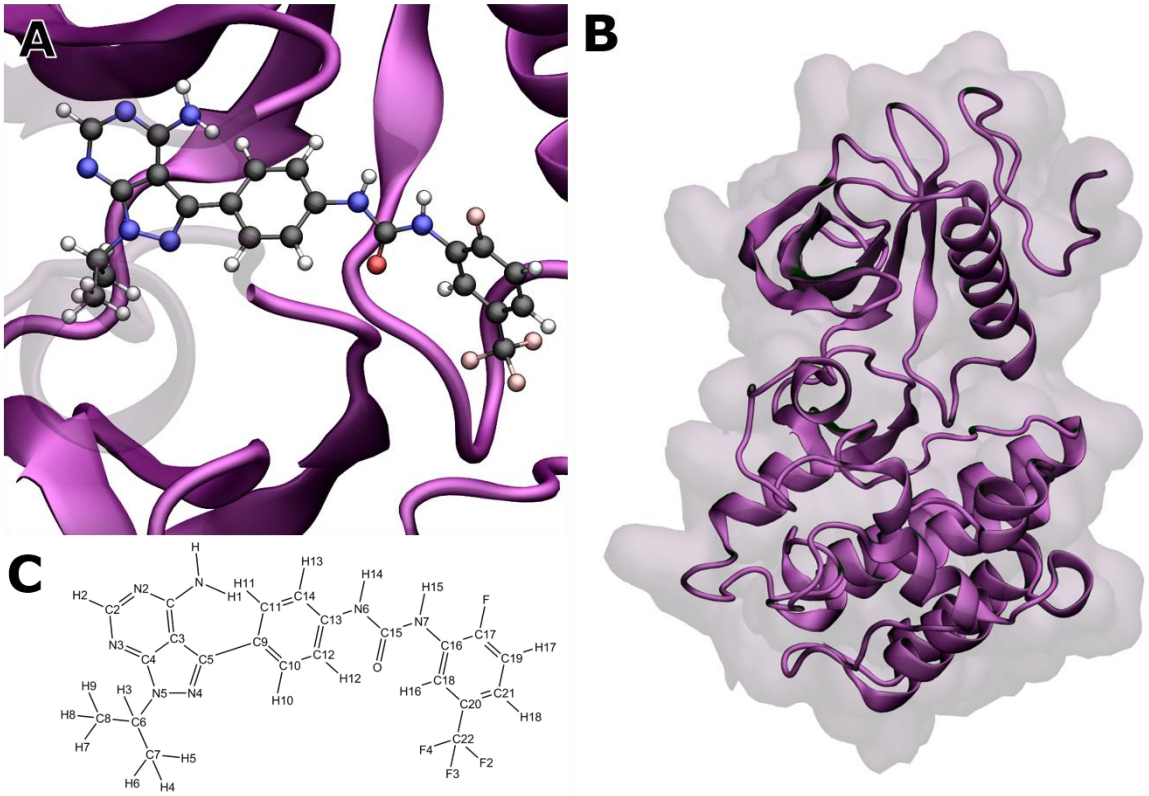

Figure 10: (A) The binding motif of AD80 in RET after MD refinement and minimization. These coordinates are used for all following calculations. (B) The RET protein after MD refinement and minimization shows a typical kinase structure. (C) Chemical structure of AD80

after which a unrestrained equilibration run of 20 ns in the *NpT* ensemble was done.[206] The last frame of this simulation was energy minimized with SANDER[192] and the analytically linearized Poisson Boltzmann (ALPB)[212] model as implicit solvent (see Figure 10 B). The obtained structure was then used for all further 1D/3D RISM-*uv* and 3D RISM-*uu* calculations and contained 4752 atoms of the protein and 53 atoms of the ligand.

The "tRNA guanine transglycosylase" (TGT) complex was modelled in a similar manner. Therefore, the structure was acquired from the pdb (pdb: 1S38) and the first processing step included the modelling of missing residues with the Modeller[207] software. After that the complex was parametrized with ff14SB force field from Amber14[213] for the protein and GAFF 1.5[208, 209] for the ligand 2-amino-8-methylquinazolin-4(3h)-one (see Figure 11 D). The zinc cofactor present in the protein was parametrized with the values[214] deposited in the ff14SB force field. Partial charges for the ligand (aminoquinazolin derivative) were calculated using the sqm tool in Amber12[192] and the AM1-BCC[210, 211] charge model. The system was then solvated in 26394 TIP3P[142] water molecules in a cubic box of the dimensions 96 Å · 94 Å · 109 Å. Then the system was subjected to local minimization, which was followed by a restrained simulation of 5 ns length with a force constant of 4 kcal/(mol · Å) applied to the C^{α} atoms of the protein in the *NVT* ensemble. Afterwards, a 5 ns long simulation with the same restraints in the *NpT* ensemble was employed. Subsequently, an unrestraint simulation of 30 ns length was done to equilibrate the system. All simulations were run at a temperature of 298.15 K via the Langevin

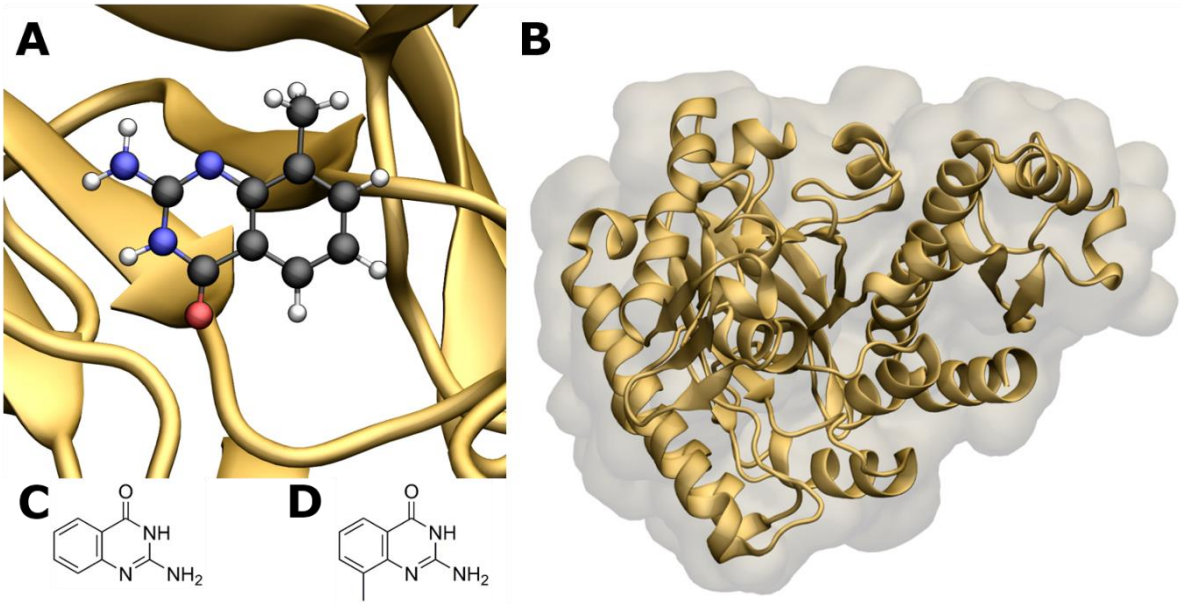

Figure 11: (A) Binding site of TGT^{MD} with bound ligand $amq^{MD, CH3}$. (B) Protein structure after MD refinement and minimization. (C) Chemical structure of 2-aminoquinazolin-4(3H)-one (amq^{H}). (D) Chemical structure of 2-amino-8-methylquinazolin-4(3h)-one ($amq^{MD, CH3}$, amq^{CH3}).

method, a pressure of 1 bar controlled by the Langevin-piston method, and used the particle mesh Ewald[215, 216] (default settings) method for the treatment of long range electrostatics. The Settle[217] algorithm was used to constrain the hydrogen atoms, and the simulations were run with a 2 fs time step in NAMD.[193] The last frame of the unrestraint trajectory was subjected to energy minimization with SANDER[192] and the ALPB[212] implicit solvent model (see Figure 11 A, B). The resulting structures for the protein and the ligand were used for further 1D RISM-*uv* and 3D RISM-*uv*/*uu* calculations. The structures are abbreviated in the following as TGT^{MD} for the protein and $amq^{MD, CH3}$ for the ligand.

The crystal structure of 1S38[218] in complex with 2-amino-8-methylquinazolin-4(3h)-one was also parametrized with the ff14SB force field of Amber14[213] for the protein and GAFF 1.5[208, 209] for the ligand as deposited in the PDB with no further refinement or modelling steps. These structures are called TGT^{CH3} for the protein and amq^{CH3} for the ligand. Furthermore the crystal structure of 1S39[218] was also downloaded from the PDB and parametrized with ff14SB from Amber14[213] for the protein and GAFF 1.5[208, 209] for the ligand 2-aminoquinazolin-4(3H)-one (see Figure 11 C). In the following text these structures for the protein are called TGT^{H} and amq^{H} for the ligand. All structures used in this chapter can be found in the electronic appendix of this work.

4.2.2 RISM-*uv* calculations

As basis for all following RISM calculations, the χ-function (result of 1D RISM-*vv*) was calculated with the dielectrically consistent (DRISM/HNC) theory[190, 191] for pure water (modified TIP3P, see chapter 3). This calculation was performed on a logarithmic grid with 512 grid points ranging from $5.98 \cdot 10^{-3}$ Å to 164.02 Å. The solvent density was set to 0.0333295 $Å^{-3}$, the temperature to 298.15 K, and the dielectric constant to 78.4. As convergence criterion, the residual norm of the direct correlation functions was set to 10^{-7}. For all necessary 1D RISM-*uv* calculations the same parameters as in the 1D RISM-*vv* case were chosen except for the convergence criterion, which was set to 10^{-5} for the maximum residual in the direct correlation functions and the number of "direct inversion of iterative subspace" (DIIS) vectors which was set to 5.

For the comparison of the closure effect on the calculated PMF and FED values, the needed 1D RISM-*uv* calculations were performed with the PSE closures of order 1-4. The 3D

RISM-*uv* calculations of the RET/AD80 complex were either done with a cuboid grid of size 120 · 110 · 130 and a grid spacing of 0.6 Å or a cuboid grid of size 260 · 240 · 280 and a grid spacing of 0.3 Å. Long range electrostatics were treated with the PME[215, 216] with order 8 and and short range interactions were cut at 14 Å. For all calculations monopole renormalization was used.[28, 130] The convergence criterion was set to 10^{-4} for the maximum residual norm of the direct correlation functions and in order to accelerate convergence 10 DIIS vectors were used.

For the calculations of the TGT complex system the same 1D RISM-*uv* parameters as described above were chosen. The corresponding 3D RISM-*uv* calculations were done with a cuboid grid of the size 250 · 230 · 290 and a grid spacing of 0.3 Å. The other 3D RISM-*uv* specific parameters were set to the same values as in the comparison of the closure effect. For all 3D RISM-*uv* calculations concerning the TGT complex system, the PSE2 closure was used. In case of the TGT complex system, all calculations were done for four different sets of structures: (1) the MD relaxed and minimised structures based on the PDB entry 1S38, which is called $TGT^{MD}/amq^{MD, CH3}$; (2) the crystal structure of TGT as it is deposited in the PDB called TGT/amq^{CH3}; (3) the crystal structure of the aminoquinazolin variant of the ligand deposited as

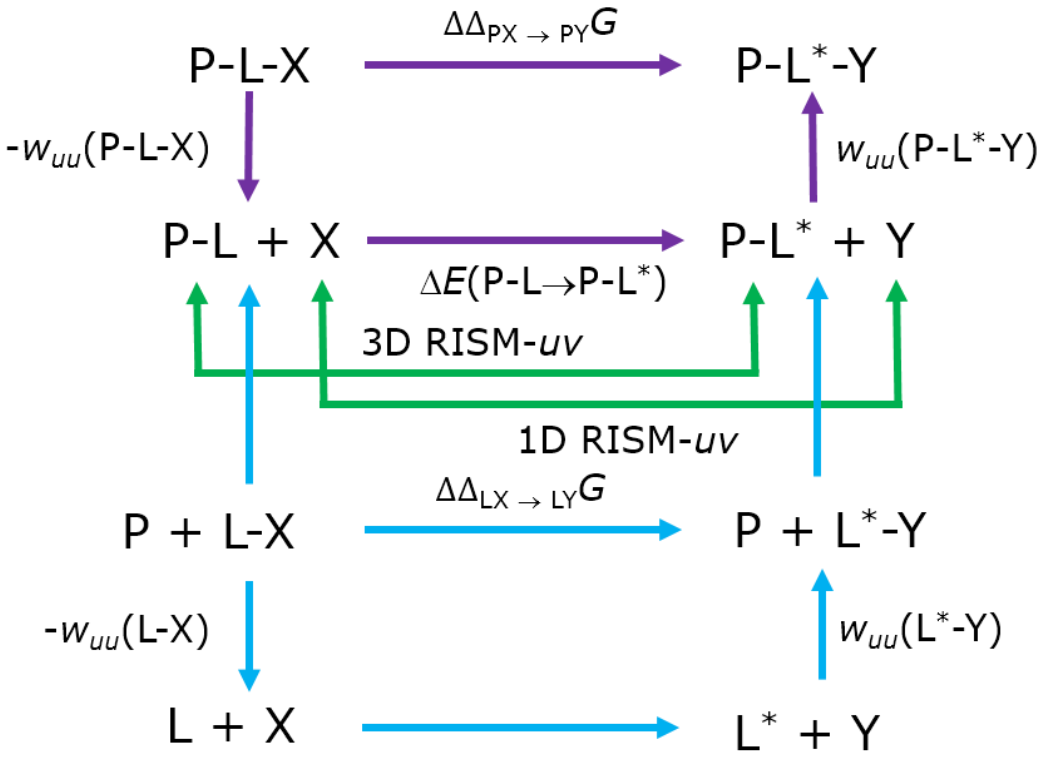

Figure 12: Exemplary thermodynamic cycle for the 3D RISM-*uu* calculations and FEDs with respect to the partial charge *q*. For all *apo* calculations only the purple circle has to be evaluated, P-L gets then substituted by P and there is no P-L*. All other contributions vanish or cancel themselves out. For the partial *holo* calculations the whole circle has to be evaluated. Note that the contribution of the protein alone vanishes in the case of the partial *holo* calculations. The green arrows only show the theory that used to calculate the given process.

found in 1S39 in the PDB, which is called TGT/amq^{H}; and (4) the crystal structure of TGT as it is deposited in the PDB, where the methyl group of the ligand is replaced by a united atom approximation called TGT/amq^{CU}. For the united atom variant σ was set to 3.905 Å, ε to 0.175 kcal mol^{-1}, and the partial charges were calculated by summing the partial charges of the original methyl atoms yielding 0.0993 e.

4.2.3 RISM-*uu* calculations

All RISM-*uu* calculations were done using the aforementioned 1D/3D RISM-*uv* calculations as input and using the same thermodynamic variables. The derivatives were calculated for the three force field parameters σ, ε, and the partial charge q. The stepsize for numerical evaluation of the derivative with a 5-point stencil was set to $\Delta = 0.05$ {Å, kcal mol^{-1}, e} around the original parameter for the atom in question throughout all calculations. Two calculation schemes were employed to derive atomwise values for either the PMF or the FED. The first scheme, is called *apo* from this point on, places every ligand atom in the "empty" binding site of the protein at the coordinates of the original protein ligand complex, which is done in a successive manner. In practical terms this means that one 3D RISM-*uv* calculation for the protein and $n \cdot 5$ 1D RISM-*uv* calculations, if n is the number of ligand atoms, were done. In the *apo* case the change from P-X to P-Y can be computed directly and in a straightforward way (see Figure 12).

The second, more elaborate, scheme, called partial *holo*, places every ligand atom into a supermolecule, either consisting of the protein with the remaining ligand atoms or the remaining ligand. The partial *holo* calculations also required that 1-3 non-bonded interactions had to be excluded and 1-4 non-bonded interactions had to be scaled by 0.5, all according to the definition of the `Amber`[192] force field. This ansatz also introduced the problem of charge neutrality for the respective ligand, and was accounted for by distributing the remaining charges onto the neighbouring atoms, as defined by the connectivity, of the ligand atom in question and calculating the electrostatic potential difference generated by introduction of this artificial charge. Practically this means that for every sampling point done for the numerical derivative two 3D RISM-*uv*, one for the partial *holo* complex and one for the partial ligand, and one 1D RISM-*uv* calculation is needed. This sums up to $5 \cdot (2n + n)$ calculations for n ligand atoms. In

the partial *holo* case one PMF point for the protein ligand complex is thus calculated in the following way:

$$\Delta\Delta G_{\mathrm{PX}\to\mathrm{PY}} = -w_{uu}(\mathrm{P}\text{-}\mathrm{L}\text{-}\mathrm{X}) + w_{uu}(\mathrm{P}\text{-}\mathrm{L}^{*} - Y) + \Delta U^{\mathrm{elec}}(\mathrm{P}-\mathrm{L}\to\mathrm{P}-\mathrm{L}^{*}), \qquad (73)$$

where $\Delta\Delta G_{\mathrm{PX}\to\mathrm{PY}}$ is the free energy change obtained from changing atom X into Y, $w_{uu}(\mathrm{P}\text{-}\mathrm{L}\text{-}\mathrm{X})$ and $w_{uu}(\mathrm{P}\text{-}\mathrm{L}^{*}-Y)$ is the PMF of the respective system. The last term $\Delta U^{\mathrm{elec}}(\mathrm{P}-\mathrm{L}\to\mathrm{P}-\mathrm{L})$ is only necessary for calculations that involve a change in the charges and guarantees that charge neutrality is maintained by:

$$\Delta U^{\mathrm{elec}} = U^{\mathrm{elec}}(\mathrm{P}-\mathrm{L},\mathrm{X}) - U^{\mathrm{elec}}(\mathrm{P}-\mathrm{L}^{*},\mathrm{Y}), \qquad (74)$$

where $U^{\mathrm{elec}}(\mathrm{P}-\mathrm{L},\mathrm{X})$ is the Coulomb potential between atom X and the partial *holo* complex and $U^{\mathrm{elec}}(\mathrm{P}-\mathrm{L}^{*},\mathrm{Y})$ is the Coulomb potential between the varied atom X and the varied partial *holo* complex. It is easily seen that ΔU^{elec} is zero if the partial charges are not varied. In addition to that the following term has also be evaluated:

$$\Delta\Delta G_{\mathrm{LX}\to\mathrm{LY}} = -w_{uu}(\mathrm{L}\text{-}\mathrm{X}) + w_{uu}(\mathrm{L}^{*} - Y) + \Delta U^{\mathrm{elec}}(\mathrm{L}\to\mathrm{L}^{*}). \qquad (75)$$

In summary this leads to the following expression for one PMF point:

$$\Delta G_{\mathrm{P\text{-}L\text{-}Y}} - \Delta G_{\mathrm{P\text{-}L\text{-}X}} = \Delta\Delta G_{\mathrm{PX}\to\mathrm{PY}} - \Delta\Delta G_{\mathrm{LX}\to\mathrm{LY}}. \qquad (76)$$

4.3 Results and discussion

4.3.1 Effect of grid sizes and PSE closure order

To evaluate the effect of the grid spacing on the PMF values, that are calculated by 3D RISM-*uu* a comparison of a high resolution grid (with a grid spacing of 0.3 Å) and a low resolution grid (with a grid spacing of 0.6 Å) are compared. In Figure 13, the differences of the atomwise PMF values between the high and low resolution grid for the RET^{apo}/AD80 system

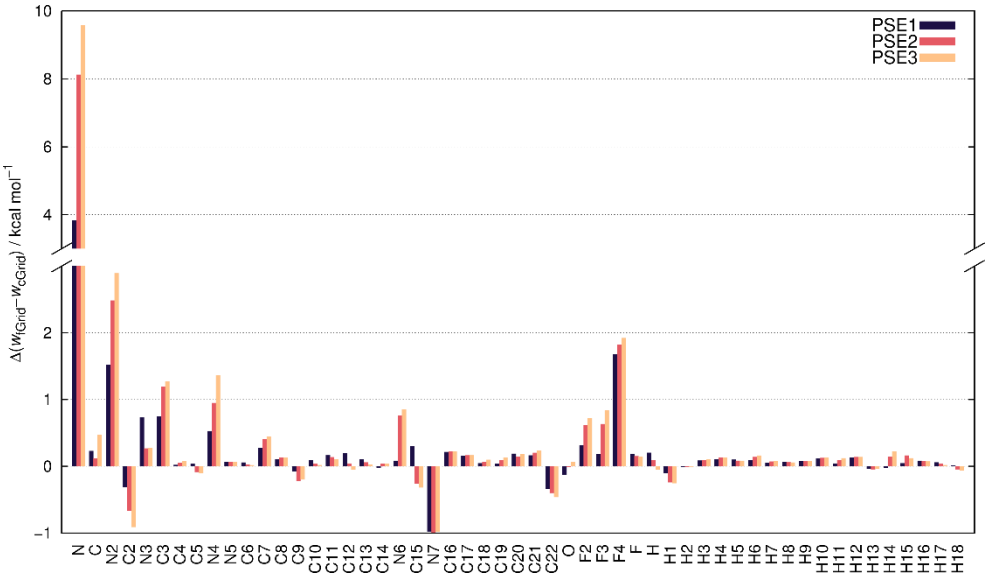

Figure 14: Differences of the PMF for the RET$^{part\ holo}$/AD80 system calculated on a higher resolution grid (w^{fGrid}, spacing of 0.3 Å) and lower resolution grid (w^{cGrid}, spacing of 0.6 Å). Data is shown for the three closures PSE1 to PSE3. The ligand atoms were placed consecutively in the partial *holo* binding site.

are shown. The range of the differences is roughly 1 kcal mol^{-1}, which shows that for this particular system the low resolution grid would be sufficient for most tasks. The oxygen atom (see Figure 10 C) shows the biggest difference between the low and high resolution grid. This could be due to a grid artifact which can arise when the center of an atom is in the direct vicinity of a grid point. All other differences are in the range of -0.2 kcal mol^{-1} to +0.4 kcal mol^{-1} which can be solely attributed to the different grid resolutions.

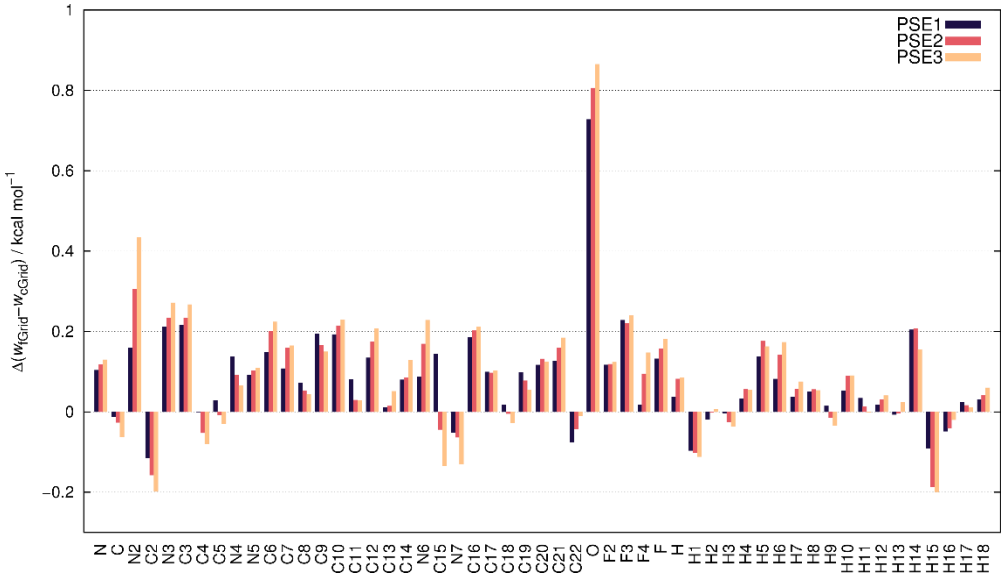

Figure 13: Differences of the PMF for the RET^{apo}/AD80 system calculated on a higher resolution grid (w^{fGrid}, spacing of 0.3 Å) and lower resolution grid (w^{cGrid}, spacing of 0.6 Å). Data is shown for the three closures PSE1 to PSE3. The ligand atoms were placed consecutively in the *apo* binding site.

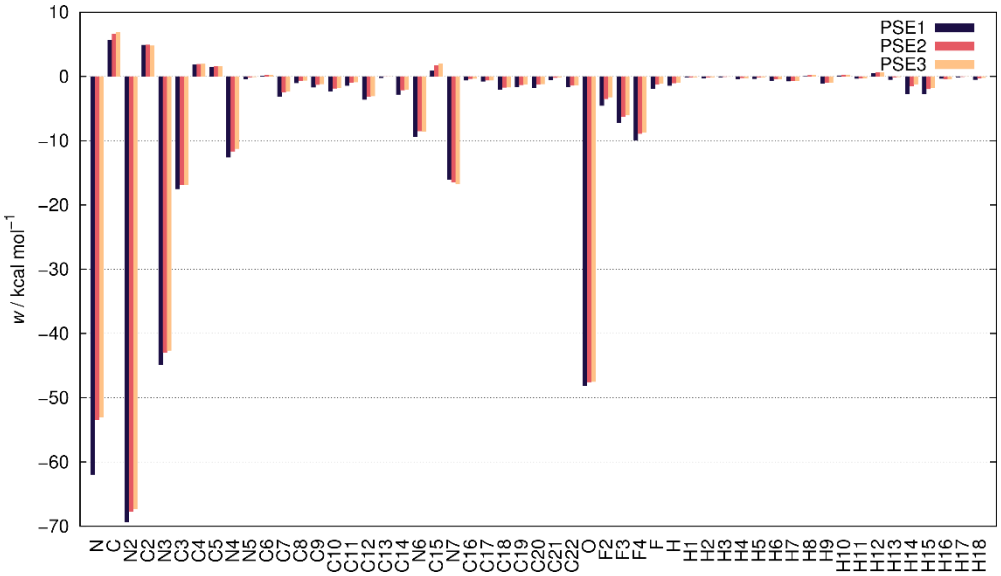

Figure 15: PMF values calculated on the w^{fGrid} for the RET$^{part\ holo}$/AD80 system and ascending PSE closure order.

When the ligand atoms are placed in the partial *holo* binding site the picture changes a little bit, as seen in Figure 14. All but two of the differences are in the range of -1 kcal mol^{-1} to +1 kcal mol^{-1} and therefore in an acceptable range. The differences of the nitrogen (N in Figure 14) are between ~(4 – 10) kcal mol^{-1}, depending on the closure relation which warrants further investigation. The working theory would be that the difference between both grid sizes is governed by the desolvation penalty which is based on the 3D RISM-*uv* calculations, where the nitrogen is shielded from the solvent environment, which could explain the significant

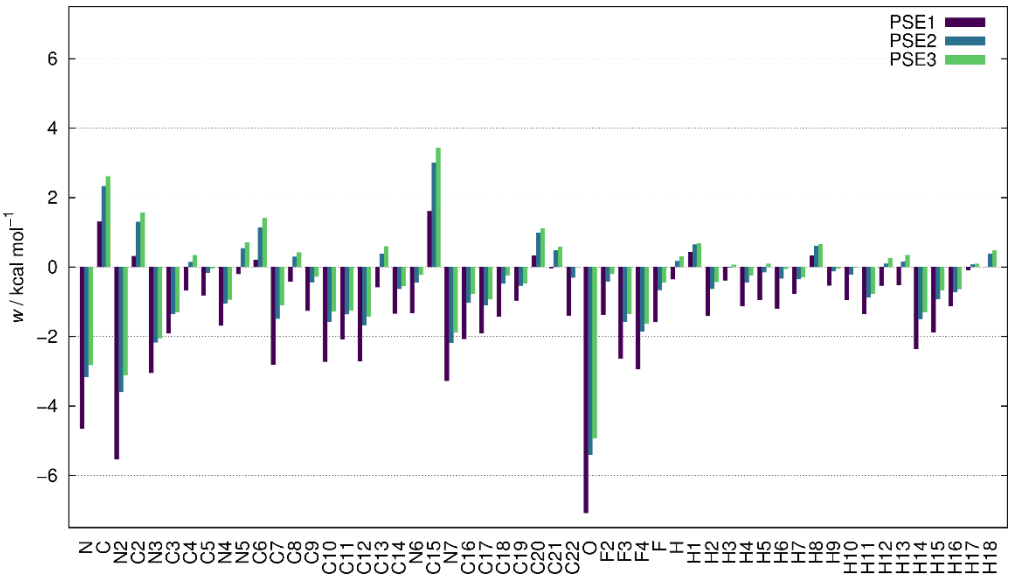

Figure 16: PMF values calculated on the w^{fGrid} for the RET^{apo}/AD80 complex and ascending PSE closure order.

difference. This implies that for the partial *holo* calculations the fine grid resolution of 0.3 seems to be better suited to yield consistent results.

The next step involves the comparison of the influence of the closure order. For this comparison only w^{fGrid} are described here. In Figure 16 the w^{fGrid} values for the *apo* binding site are shown for ascending PSE closure order. In general the PSE1 closure shows significant deviations from the other closures and in particular the PSE3 closure. Taking into consideration the working hypothesis that the PMF values calculated with a higher order PSE are superior to those that are calculated with a lower order PSE, which is not completely unwarranted (see Ref. [219]). Following this line of argument the calculations with the PSE3 order were flagged as the "gold standard" and the deviation between either PSE1 or PSE2 was calculated. If PSE1 and PSE3 are compared, the standard deviation of the differences is σ = 3.12 kcal mol^{-1} and the Pearson correlation coefficient (used here to reveal possible anti-correlation) is R = 0.96. The same comparison done for PSE2 and PSE3 shows a standard deviation of the differences of only σ = 0.73 kcal mol^{-1} with a correlation coefficient of R = 0.99. It seems to be the case that for the calculations in the *apo* binding site the PSE2 closure delivers the best compromise between speed, convergence behaviour, and accuracy.

In the case of the partial *holo* complex the picture is comparable to the *apo* case as seen in Figure 15. The standard deviation of the differences between the calculations with PSE1/PSE3 is σ = 8.8 kcal mol^{-1} and for PSE2/PSE3 σ = 0.93 kcal mol^{-1} respectively. The Pearson correlation coefficients for both PSE1/PSE3 and PSE2/PSE3 is R = 0.99. This suggests that the closure has a more pronounced effect on partial *holo* complex calculations than on *apo* complex calculations. The conclusion would be to use, as in the *apo* case PSE2 as a standard because of same reasons.

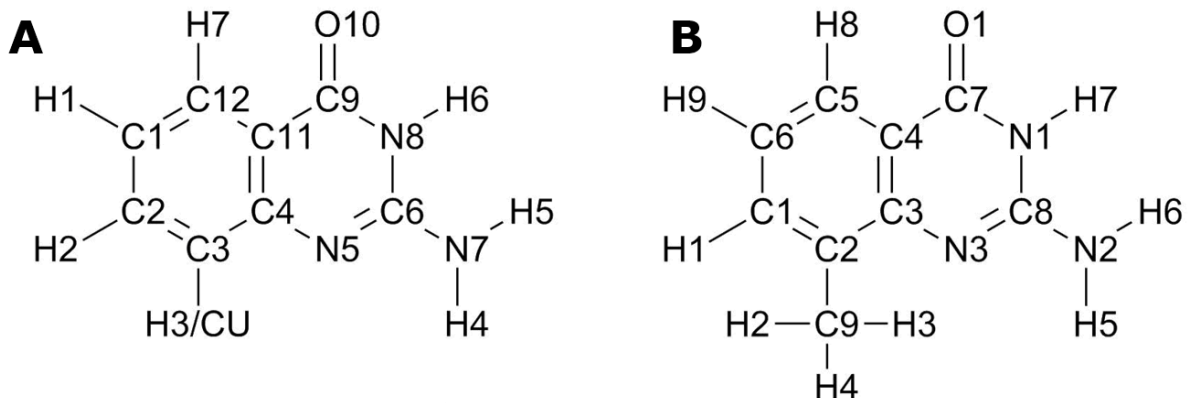

Figure 17: (A) Atom assignment for the amq^{H} ligand. For further investigation of geometric effects on the partial *holo* calculations. (B) Atom assignment for the amq^{CH3} and amq^{MD,CH3} ligand.

4.3.2 A tool for rational drug design: a case study

In this subchapter the usefulness of free energy derivatives with regard to the drug design process will be addressed. Therefore, the atomwise FEDs and PMFs for the four corresponding complex pairs $TGT^{MD,\,CH3}/amq^{CH3}$, TGT^{CH3}/amq^{CH3}, TGT^{CH3}/amq^{CU}, and TGT^{H}/amq^{H} were calculated for either the *apo* or the partial *holo* case. The FEDs were calculated with regard to the force field parameters σ, ε and q.

Table 2: Sum of the atomwise PMF (w) values for all calculation modes and complexes. For the two aminoquinazolin derivatives the experimental K_i values are also shown. All calculations were done on the finer grid (Δx = 0.3 Å) for the respective structures in their native environment.

complex	calc. mode	w_{sum}^{fGrid} (kcal/mol)	exp. K_i
$TGT^{MD,CH3}/amq^{MD,CH3}$	partial *holo*	-111.10	-
TGT^{CH3}/amq^{CH3}	partial *holo*	-115.40	7 μM[a]
TGT^{H}/amq^{H}	partial *holo*	-100.30	20 nM - 50 nM[a]
$TGT^{MD,CH3}/amq^{MD,CH3}$	*apo*	-15.14	-
TGT^{CH3}/amq^{CH3}	*apo*	-14.90	-
TGT^{CU}/amq^{CU}	*apo*	-15.75	-
TGT^{H}/amq^{H}	*apo*	-14.67	-

[a]Meyer *et al.*,[218] $\Delta\Delta G^{exp}_{CH3\rightarrow H}$=-3.47 kcal · mol^{-1} − -2.93 kcal · mol^{-1}

Firstly, it is checked if the summation of the atomwise PMF values correlates with the experimental trend and if significant changes arise between the MD relaxed and crystal structures. In Table 2, the results are summarised. In case of the partial *holo* calculations the PMF sum for the MD relaxed structure (superscript MD, CH3) and the crystal structure (superscript CH3) are -111.10 kcal mol^{-1} and -115.40 kcal mol^{-1} respectively and therefore in the same range. The same result can be observed for the *apo* calculations in which the difference between the MD relaxed structure and crystal structure shrinks to -0.24 kcal mol^{-1}, which indicates that the MD simulation is not imperative in this particular case.

These results demonstrate that the PMF sum alone is not enough to distinguish between a good and mediocre binder. Therefore they are crucial for the next chapter which investigates if *apo* PMF calculations can be used to define a novel scoring function. Additionally, the partial *holo* calculations seem to suffer from a conceptional shortcoming in the form of the fixed ligand geometry. To investigate this in detail the separate contributions to the PMF for the partial *holo* (only the complex part) and *apo* calculations are shown in Table 3. In the partial *holo* case the

desolvation penalty values are generally higher than those in the *apo* case and the average difference between the potential part and the desolvation penalty is 9.91 kcal mol^{-1}, in contrast to only 2.25 kcal mol^{-1} in the *apo* calculations. For the partial *holo* calculations none of the desolvation penalties has a negative sign instead of the *apo* calculations where the solvation process seems to be beneficial in some cases.

Table 3: Contributions to the complex PMF for amq[H] ligand atoms (in the crystal structure) in the partial *holo* and *apo* case. The PMF is split into the potential (u) and desolvation penalties (w^v).

	partial *holo*		*apo*	
atom	u (kcal mol^{-1})	w^v (kcal mol^{-1})	u (kcal mol^{-1})	w^v (kcal mol^{-1})
C1	-1.75	14.79	-2.26	-0.06
C2	-1.22	11.80	-1.93	-0.23
C3	-1.43	14.13	-1.98	-1.23
C4	-0.65	24.47	-2.35	-0.98
N5	-0.68	70.03	0.14	2.41
C6	-2.64	95.41	-2.85	-3.03
N7	0.01	118.77	0.97	5.54
N8	-2.04	77.44	-2.27	-0.73
C9	-2.72	78.33	-2.83	-4.23
O10	-1.81	80.08	-1.27	2.25
C11	-0.70	24.39	-2.20	-1.84
C12	-1.40	10.99	-2.41	-1.39
H1	1.29	7.53	1.32	0.91
H2	0.48	6.29	0.51	0.29
H3	-0.51	6.12	-0.49	-1.03
H4	-1.34	19.04	-0.98	2.18
H5	-1.33	19.56	-1.08	0.94
H6	-1.33	8.43	-0.78	0.49
H7	-0.35	5.09	-0.42	-1.20

From this it is obvious that one has to be cautious with the results of the partial *holo* calculations, because they seem to be dominated by the contributions of the desolvation penalty. A closer look reveals that the biggest differences, 28.85 kcal mol^{-1} for N7, can be seen for atoms that are surrounded by neighbouring ligand atoms (see Figure 17). This seems to be the correct physical description because placing an atom in such a confined space will naturally be penalized. To further investigate this theory isosurfaces of the *g*-function of a 3D RISM-*uv* calculation are

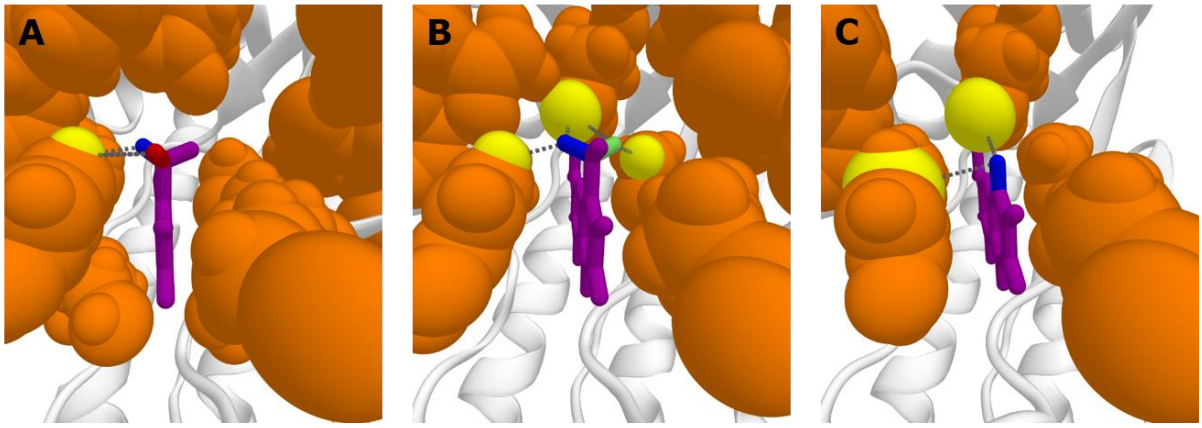

Figure 19: Binding modes of the three complex structures with all residues of the protein residues in a 5 Å radius in orange. (A) shows the binding mode of the $TGT^{MD}/amq^{MD,CH3}$ complex: distance of H3 (blue) to the closest protein atom (yellow, HD2 of Tyr 106) is 2.35 Å and of H2 (red) to the same atom is 3.14 Å. (B) shows the binding mode of the TGT/amq^{CH3} complex: distance of H3 (blue) to the HD2 atom of Tyr 106 (left side) is 2.30 Å and to the OD1 atom of Asp 102 (buried in the binding site) is 2.52 Å. For H4 (green) the distance to the OD1 atom of Asp 102 (buried in the binding site) is 2.81 Å and 2.46 Å to the HE3 atom of Met 260 (right side). (C) shows the binding mode of the complex TGT/amq^{H}: the distance of the H3 atom (blue) to the CD2 atom of Tyr 106 is 3.76 Å and 3.14 to the OD1 atom of Asp 102.

shown in Figure 18. At the position of atom N7 Figure 18 shows no water density found for an isovalue of 2. This naturally leads to a strong penalisation in the resulting PMF.

Now the focus changes to the atomwise calculation of the free energy derivatives for the four TGT complexes and in particular the TGT^{CU}/amq^{CU} (CU being an abbreviation for the united atom variant) and the TGT^{H}/amq^{H} complex. For all other complexes the corresponding

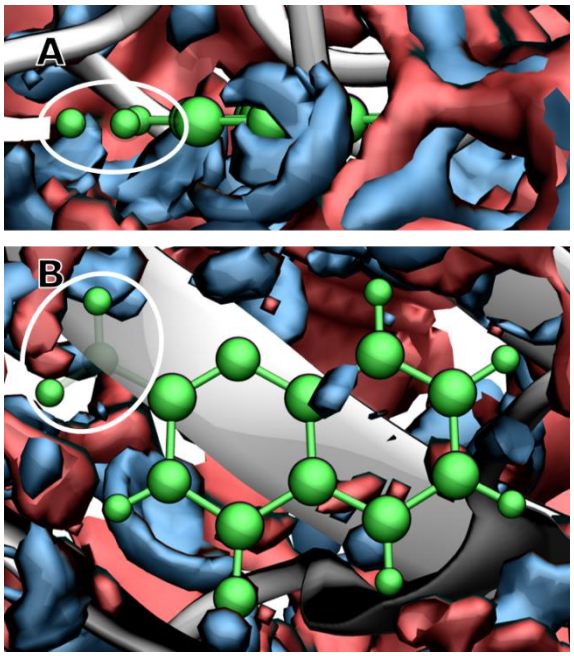

Figure 18: A) Shows the partial amq^{H} ligand (N7 is missing) in a top view inside the binding site. The white circle highlights the part of interest, where no water density is found. B) Shows the partial amq^{H} ligand (N7 is missing) in a front view inside the binding site. The white circle highlights the part of the molecule where no water density is found. The oxygen densities are shown in red and the hydrogen densities in blue. Isosurfaces are shown for a isovalue of 2.

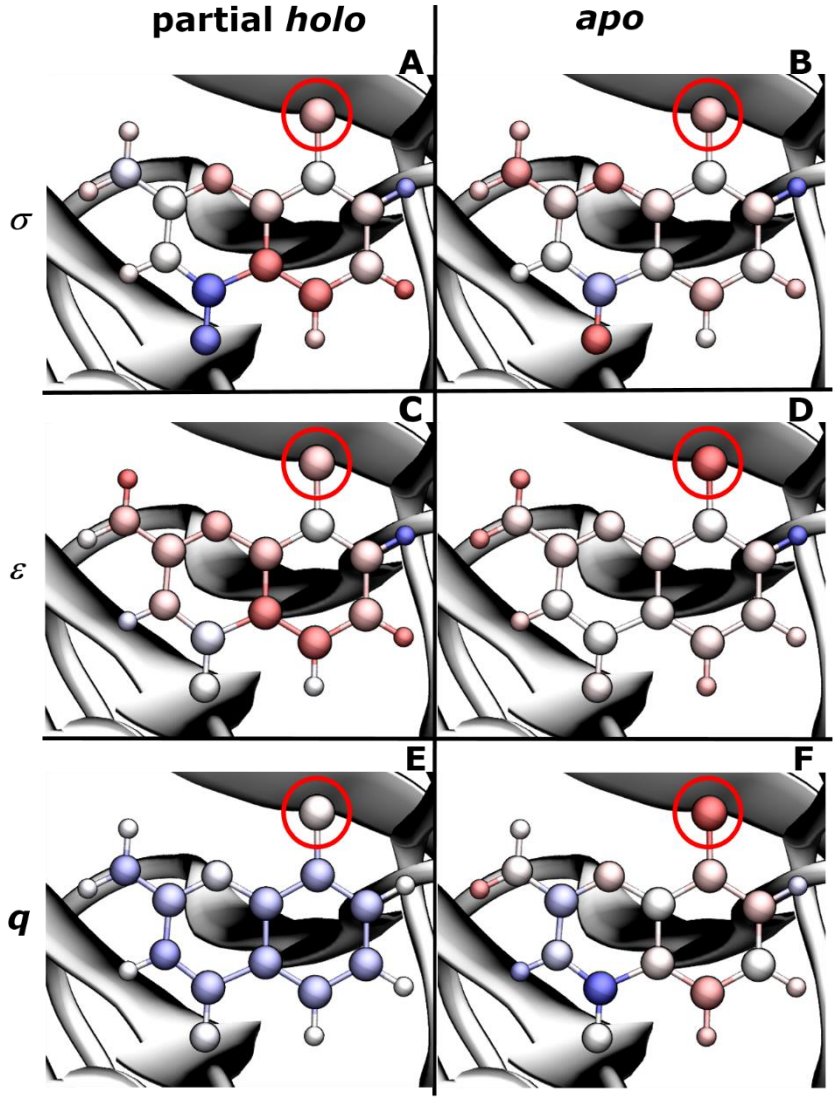

Figure 20: Atomwise FEDs for the TGT^{CU}/amq^{CU} system. The upper row shows the FEDs with respect to the σ value for the partial *holo* (A) and *apo* (B) calculations. In the middle row the FEDs with respect to the Lennard-Jones Parameter ε are shown for the partial *holo* (C) and *apo* (D) are shown. The last row shows the FEDs in regard to the partial charge q for the partial *holo* (E) and *apo* (F) calculations. The atoms are colour coded from red to white up to blue with red associated with a negative FED value (means that the parameter has to get smaller to approach an optimum) and blue with a positive FED value (means that the parameter has to get bigger to approach an optimum). The atom group of particular interest is encircled in red.

FEDs are shown in the appendix. Starting with the derivative with respect to the partial charge for which the resulting FEDs are shown in Figure 20, Figure 21, and the corresponding binding modes of the ligands in Figure 19. The FEDs are shown in a colour scale from red to white up to blue which represents negative, optimal and positive FED values respectively. The atom group of interest, consisting of CU (united atom methyl group) or H3, the partial *holo* FEDs with respect to the charge (Figure 20 (E, F) are positive (Table 4 upper part). This means that a less negatively charged group would lead to a better binder. The *apo* calculations show that the derivative is negative for TGT^{CU}/amq^{CU} and slightly positive for TGT^{H}/amq^{H}. To further assess these results a look at the actual charges for the amq^{H} and amq^{CU} ligand is helpful. Both ligands

have a positive partial charge, but the amq^{H} ligand is less negatively charged with a partial charge of 0.15 e, that leads to a charge difference of Δq = -0.06 e ($q(amq^{H}) - q(amq^{CU})$). With this information it is now possible to get a measure of how large the influence of the actual change of the charges from CH3 to H could be by calculating $\partial w/\partial q \cdot \Delta q$, which is a linear approximation. This leads to a change of 1.93 kcal mol^{-1} for the partial *holo* calculation directly at the hotspot, and interestingly to a change of -8.08 kcal mol^{-1} at the neighbouring C3 atom, which is a member of the ring system. To account for a possible hysteresis effect the linear approximation for $-\partial w/\partial q \cdot \Delta q$ was also calculated and shows the opposite trends. Despite this being the desired outcome (the trend gets inverted) it also shows that there exists a hysteresis effect, in part caused by the rigid structures used for the calculations, because the effects do not cancel out completely (4.29 kcal mol^{-1}).

For the *apo* calculation the charge difference leads to a change of -0.19 kcal mol^{-1} directly at the position of H3 but to a change of 0.08 kcal mol^{-1} at the neighbouring ring position. The control calculation of $-\partial w/\partial q \cdot \Delta q$ shows that directly at the hot spot the free energy change upon introduction of CH3 group would be slightly negative (-0.01 kcal mol^{-1}). At the neighbouring C3 position the same calculation shows that the effect would be more or less cancelled out. This shows that for the *apo* calculations the hysteresis effect is larger than for the partial *holo* calculations. All other atoms have rather minor contributions compared to that and cancel each other more or less out.

What is interesting about this is that both calculation methods would lead to a total free energy change, by summing all contributions, from amq^{CU} to amq^{H} of -5.75 kcal mol^{-1} for the partial *holo* calculation and -0.12 kcal mol^{-1} for the *apo* calculation. This means that both calculation methods give the same general trend but show different signs at the hotspot position. This interesting effect should be further investigated. It is also of note that both methods seem to be able, although with opposing trends at the same site, to account for the change in the partial charges of the whole molecule, and it is reassuring to see that the FEDs are able to resolve that local changes can have a notable effects at another position.

So far, the change introduced by the replacement of the methyl group by a hydrogen points, according to the calculated FEDs, in the right direction. But amq^{H} is probably not the ideal ligand and maybe it is possible to get a better binding ligand by the introduction of a nitrogen at the position of C3, which would introduce an even less negative charge at the position H3/CU.

Table 4: FEDs with respect to q, σ, and ε for TGT^{CU}/amq^{CU} (denoted with superscript 1) and TGT^{H}/amq^{H} (denoted with superscript 0). The differences $\Delta\{q, \sigma, \varepsilon\}$ are always calculated by $\Delta\{q, \sigma, \varepsilon\} = \Delta\{q, \sigma, \varepsilon\}^{(0)} - \Delta\{q, \sigma, \varepsilon\}^{(1)}$.

				apo				partial *holo*			
atoms	$q^{(0)}$	$q^{(1)}$	$\Delta q^{(0)-(1)}$	$\partial w/\partial q^{(1)}$	$\partial w/\partial q^{(0)}$	$\partial w/\partial q^{(1)}\cdot\Delta q$	$-\partial w/\partial q^{(0)}\cdot\Delta q$	$\partial w/\partial q^{(1)}$	$\partial w/\partial q^{(0)}$	$\partial w/\partial q^{(1)}\cdot\Delta q$	$-\partial w/\partial q^{(0)}\cdot\Delta q$
C1	-0.165	-0.159	-0.006	-0.07	-0.89	0.00	-0.01	138.88	53.61	-0.83	0.32
C2	-0.087	-0.092	0.005	-1.58	-0.40	-0.01	0.00	136.67	50.67	0.68	-0.25
C3	**-0.136**	**-0.076**	**-0.060**	**-1.41**	**-1.05**	**0.08**	**-0.06**	**135.26**	**48.15**	**-8.08**	**2.87**
C4	0.213	0.210	0.003	-0.22	0.38	0.00	0.00	137.32	50.33	0.41	-0.15
N5	-0.681	-0.680	-0.001	-1.23	1.90	0.00	0.00	100.20	35.94	-0.09	0.03
C6	0.667	0.665	0.002	2.84	3.78	0.01	-0.01	136.55	49.75	0.26	-0.09
N7	-0.888	-0.887	-0.001	-0.70	1.71	0.00	0.00	149.17	67.99	-0.13	0.06
N8	-0.507	-0.506	-0.001	2.09	1.60	0.00	0.00	155.73	66.19	-0.16	0.07
C9	0.719	0.719	0.000	6.12	4.06	0.00	0.00	140.41	53.03	-0.01	0.01
O10	-0.624	-0.624	0.000	-0.45	-0.58	0.00	0.00	96.67	70.53	0.00	0.00
C11	-0.210	-0.204	-0.006	-0.85	-0.68	0.01	0.00	141.80	53.32	-0.85	0.32
C12	-0.052	-0.058	0.006	-2.42	-1.48	-0.01	0.01	138.64	51.66	0.83	-0.31
H1	0.139	0.139	0.000	-1.29	2.29	0.00	0.00	46.80	16.54	0.00	0.00
H2	0.136	0.136	0.000	2.32	2.33	0.00	0.00	40.95	14.01	0.00	0.00
H3/CU	**0.151**	**0.099**	**0.052**	**-3.70**	**0.24**	**-0.19**	**-0.01**	**37.35**	**22.21**	**1.93**	**-1.15**
H4	0.415	0.413	0.002	-0.68	-3.30	0.00	0.00	92.57	87.52	0.14	-0.13
H5	0.415	0.413	0.002	-2.83	-0.59	0.00	0.00	93.95	87.90	0.14	-0.13
H6	0.336	0.336	0.000	3.96	-1.72	0.00	0.00	94.33	63.08	0.00	0.00
H7	0.156	0.156	0.000	-1.97	5.23	0.00	0.00	43.68	14.95	0.00	0.00
net effect						-0.12	-0.07			-5.75	1.46
hysteresis avg.						-0.02				-3.60	
	$\sigma^{(0)}$	$\sigma^{(1)}$	$\Delta\sigma^{(0)-(1)}$	$\partial w/\partial\sigma^{(1)}$	$\partial w/\partial\sigma^{(0)}$	$\partial w/\partial\sigma^{(1)}\cdot\Delta\sigma$	$-\partial w/\partial\sigma^{(0)}\cdot\Delta\sigma$	$\partial w/\partial\sigma^{(1)}$	$\partial w/\partial\sigma^{(0)}$	$\partial w/\partial\sigma^{(1)}\cdot\Delta\sigma$	$-\partial w/\partial\sigma^{(0)}\cdot\Delta\sigma$
C1	3.400	3.400	-	-0.71	-2.38	-	-	-0.61	-1.15	-	-
C2	3.400	3.400	-	-1.30	-1.39	-	-	-0.63	-0.62	-	-
C3	3.400	3.400	-	-0.45	-1.51	-	-	0.08	-1.17	-	-
C4	3.400	3.400	-	-1.24	-1.54	-	-	-0.73	-0.99	-	-
N5	3.250	3.250	-	-3.26	3.11	-	-	-1.26	4.41	-	-
C6	3.400	3.400	-	-1.19	-1.48	-	-	-0.03	-0.60	-	-
N7	3.250	3.250	-	-3.31	1.93	-	-	0.81	7.41	-	-
N8	3.250	3.250	-	-0.35	-0.94	-	-	-0.27	-1.13	-	-
C9	3.400	3.400	-	1.33	1.22	-	-	2.89	1.93	-	-
O10	2.960	2.960	-	-4.28	-2.57	-	-	2.53	2.77	-	-
C11	3.400	3.400	-	-0.65	-0.91	-	-	-1.85	-1.64	-	-
C12	3.400	3.400	-	-1.94	-1.93	-	-	-1.81	-0.84	-	-
H1	2.600	2.600	-	-2.26	3.41	-	-	-2.00	3.49	-	-
H2	2.600	2.600	-	2.13	1.86	-	-	2.19	2.25	-	-
H3/CU	**2.600**	**3.905**	**-1.305**	**-2.38**	**0.05**	**3.11**	**0.07**	**-1.09**	**-0.12**	**1.43**	**-0.16**
H4	1.069	1.069	-	-2.17	-2.57	-	-	-0.64	-2.29	-	-
H5	1.069	1.069	-	-1.86	-2.27	-	-	-0.80	-2.52	-	-
H6	1.069	1.069	-	-0.30	-1.99	-	-	-0.46	-2.76	-	-
H7	2.600	2.600	-	-0.62	0.44	-	-	-0.86	0.24	-	-
net effect						3.11	0.07			1.43	-0.16
hysteresis avg.						1.52				0.79	
	$\varepsilon^{(0)}$	$\varepsilon^{(1)}$	$\Delta\varepsilon^{(0)-(1)}$	$\partial w/\partial\varepsilon^{(1)}$	$\partial w/\partial\varepsilon^{(0)}$	$\partial w/\partial\varepsilon^{(1)}\cdot\Delta\varepsilon$	$-\partial w/\partial\varepsilon^{(0)}\cdot\Delta\varepsilon$	$\partial w/\partial\varepsilon^{(1)}$	$\partial w/\partial\varepsilon^{(0)}$	$\partial w/\partial\varepsilon^{(1)}\cdot\Delta\varepsilon$	$-\partial w/\partial\varepsilon^{(0)}\cdot\Delta\varepsilon$
C1	0.598	0.598	-	-0.86	-1.48	-	-	-0.78	-1.17	-	-
C2	0.598	0.598	-	-0.96	-1.10	-	-	-0.68	-0.91	-	-
C3	0.598	0.598	-	-0.58	-1.00	-	-	-0.40	-0.98	-	-
C4	0.598	0.598	-	-0.99	-1.15	-	-	-0.80	-0.98	-	-
N5	1.181	1.181	-	-1.02	-0.02	-	-	-0.84	0.09	-	-
C6	0.598	0.598	-	-1.12	-1.29	-	-	-0.73	-1.04	-	-
N7	1.181	1.181	-	-1.12	-0.32	-	-	-0.83	0.17	-	-
N8	1.181	1.181	-	-0.53	-0.58	-	-	-0.67	-0.78	-	-
C9	0.598	0.598	-	-0.07	-0.01	-	-	0.51	0.24	-	-
O10	1.459	1.459	-	-0.77	-0.48	-	-	-0.30	-0.14	-	-
C11	0.598	0.598	-	-0.63	-0.79	-	-	-1.12	-1.14	-	-
C12	0.598	0.598	-	-1.12	-1.16	-	-	-1.07	-0.86	-	-
H1	0.104	0.104	-	-1.31	5.57	-	-	-1.13	5.63	-	-
H2	0.104	0.104	-	2.92	2.49	-	-	3.07	3.09	-	-
H3/CU	**0.104**	**1.216**	**-1.112**	**-3.56**	**-0.42**	**3.96**	**-0.47**	**-0.77**	**-0.73**	**0.85**	**-0.62**
H4	0.109	0.109	-	-3.13	-3.72	-	-	-1.20	-3.37	-	-
H5	0.109	0.109	-	-2.78	-3.27	-	-	-0.48	-3.48	-	-
H6	0.109	0.109	-	-1.61	-2.94	-	-	0.76	-3.79	-	-
H7	0.104	0.104	-	-2.05	-0.11	-	-	0.08	-0.39	-	-
net effect						3.96	-0.47			0.85	-0.62
hysteresis avg.						2.21				0.73	
sum net effect						6.95				-3.47	
sum hysteresis						3.71				-2.08	

q: e; σ: Å; ε: kcal mol^{-1}; $\partial w/\partial q$: kcal mol^{-1} · e^{-1}; $\partial w/\partial\sigma$: kcal mol^{-1} · Å^{-1}; $\partial w/\partial\{q,\sigma\varepsilon\}\cdot\Delta\{q,\sigma,\varepsilon\}$: kcal mol^{-1}; $\Delta\Delta G^{exp}_{CH3\rightarrow H}$ =-3.47 kcal mol^{-1} − -2.93 kcal mol^{-1};[218] $\Delta\Delta G^{calc,pHolo}_{CH3\rightarrow H}$ =-2.08 kcal mol^{-1}; $\Delta\Delta G^{calc,apo}_{CH3\rightarrow H}$ =3.71 kcal mol^{-1}

A look into the middle part of Table 4 reveals that the derivatives with respect to σ for the systems TGT^{CU}/amq^{CU} and TGT^{H}/amq^{H}, which are also visualised in Figure 20 and Figure 21, have both a negative sign. The effect is not that strong for the TGT^{H}/amq^{H} system where the derivative is only -0.12 kcal $mol^{-1} \cdot Å^{-1}$ and therefore is in the vicinity of minimum or maximum. This could only be decided by the calculation of the second order derivative. The *apo* calculations show a similar trend with the derivative for the TGT^{CU}/amq^{CU} system being negative and for the TGT^{H}/amq^{H} complex slightly positive. The linear approximation of the free energy change show for both calculation schemes a positive effect, with 3.11 kcal mol^{-1} and 1.43 kcal mol^{-1} respectively. Both calculation schemes show a hysteresis effect in regard to the free energy change, which is stronger for the *apo* calculation than for the partial *holo* calculation. Nonetheless both calculations schemes show that for the design choice the results of the FEDs with respect to σ could be interpreted to point into the same direction. Finding a group that shows a more positive partial charge and at the same time is bulkier than a methyl group is not easy.

Three possibilities arise: one could keep the proposed change at position C3 that resulted from the FED with respect to the charge, which showed to have a rather significant effect on the resulting energies, disregarding the information about the derivative with respect to the σ-value. Another possibility is to do the exact opposite thing and disregard the FED information with respect to the charge, which then leads to the proposition to introduce an amine or thiol group. The third possibility is to do the described change for C3 to N and at the same time keep the methyl group or exchange it by an ethyl group, to make it even more "bulky."

At last a look at the FEDs with respect to the ε-value for both systems and calculation schemes is warranted. Figure 20 and Figure 21 show the visualisation of the FEDs for both systems and the lower part of Table 4 the accompanying data. All FEDs for H3/CU are negative for both systems and calculation schemes. The calculations for the TGT^{H}/amq^{H} system show smaller derivatives than the derivatives for the TGT^{CU}/amq^{CU} system. This implies that a group with a bigger ε-parameter could have a positive effect on the binding characteristics. For both calculation methods this leads to the same trend of a positive free energy change resulting from the derivatisation of the methyl group with the hydrogen. The calculation schemes also show a hysteresis effect, with the *apo* calculations showing a stronger effect than the partial *holo* calculations. These results are in accordance to the aforementioned trends for the design direction and keep the three described changes as viable possibilities.

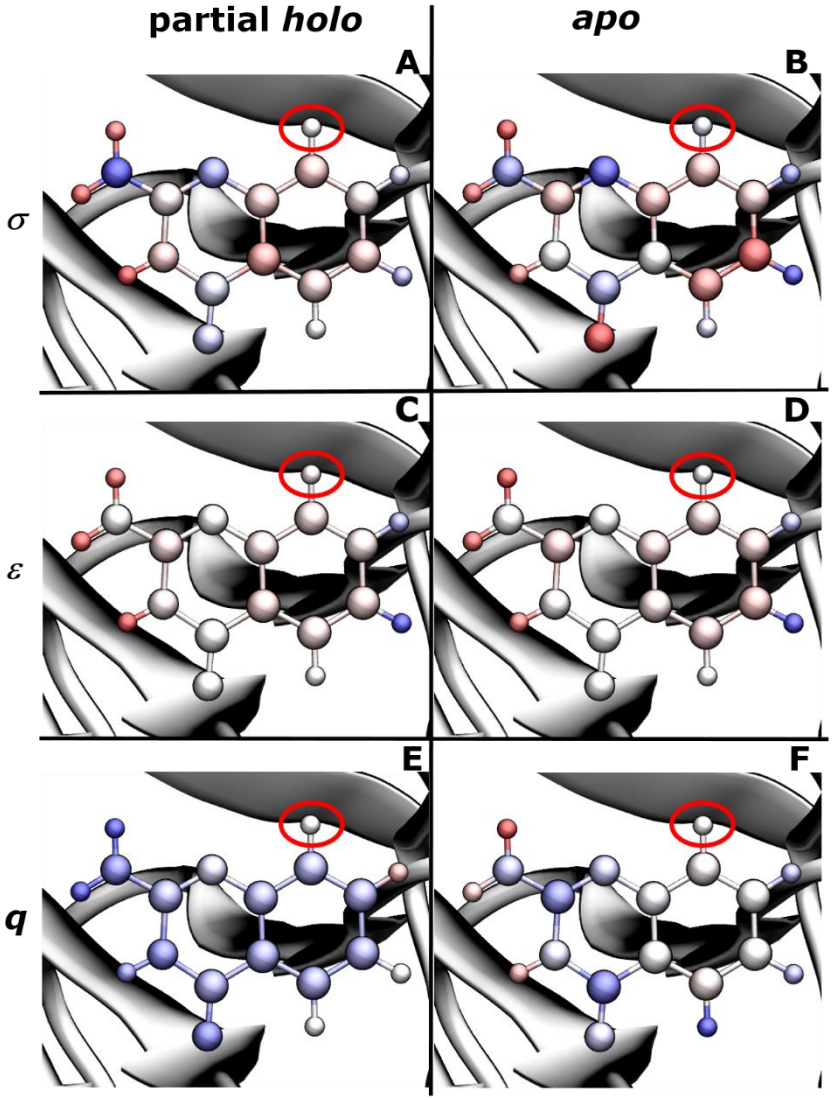

Figure 21: Atomwise FEDs for the TGT/amq^{H} system. The upper row shows the FEDs with respect to the σ value for the partial *holo* (A) and *apo* (B) calculations. In the middle row the FEDs with respect to the Lennard-Jones Parameter ε are shown for the partial *holo* (C) and *apo* (D) are shown. The last row shows the FEDs in regard to the partial charge *q* for the partial *holo* (E) and *apo* (F) calculations. The atoms are colour coded from red to white up to blue with red associated with a negative FED value (means that the parameter has to get smaller to approach an optimum) and blue with a positive FED value (means that the parameter has to get bigger to approach an optimum). The atom group of particular interest is encircled in red.

To close the examination of the hotspot FEDs up, the key conclusions and results are presented in the next paragraph. First: the calculations implicate that the biggest effect on the binding affinity, only from the standpoint of the influence the non-bonded force field parameters have, has the change in the partial charges of amq^{CU} to amq^{H}. This effect overcompensates, only for the partial *holo* calculations, the negative change from the methyl group to the hydrogen that is seen for the changes of the σ- or ε-values. If all effects of the linear approximation are summed up and the hysteresis effect is accounted for, the change from the methyl group to the hydrogen leads to a cumulative effect of 3.71 kcal mol^{-1} for the *apo* calculations and -2.08 kcal mol^{-1} for the partial *holo* calculations. Experimentally, the change in

the binding free energy amounts to -3.47 kcal mol^{-1} – -2.93 kcal mol^{-1}. It should also be noted that the *apo* calculations are not able to reproduce the right trends with regard to the hysteresis effect for the partial charge and the σ value.

For the decision, which design direction to choose, this could mean that the change of C3 to a N could have a greater effect on the binding free energy than the proposed changes that would account for the derivatives with regard to σ and ε. For the *apo* calculations the picture inverts. Here the effect that is seen for the changes for σ and ε outweigh the binding affinity that is gained through the change of the partial charges. Because it is known that amq^{H} binds better to TGT one could argue that the effect that is achieved through the changed partial charges outweighs the effects seen for σ and ε. But this remains to be shown by further analysis of the system at hand. Furthermore, at first glance these results seem to be counter intuitive, because the inspection of the binding site reveals that the amq^{CH3} ligand is tightly wrapped in the binding site and could probably benefit from a smaller ligand footprint.

Second: the optimal derivatisation for this position could be far from a hydrogen atom, if all contemplable derivatives are considered. The most promising design direction would point into the direction of a change of C3 to a N or the introduction of a ethyl group. It should be noted that a lot of effects play a role for the binding process and actual binding affinity and the results presented here should not be over-interpreted.

Third: the focus in this study was laid on the H3/CU position because experimental data for that derivatisation was already published and this rather subtle change has a quite pronounced effect on the measured binding affinities. This makes it a good test system, but there are other interesting sites in the molecule that could have an even greater effect on the binding affinity. For example: position H6 shows a rather strong derivative with respect to the partial charge, although a derivatisation at this site could be synthetically hard, introduction of a more negatively charged atom or group could have a strong effect. This shows, that a *posteriori* "prediction" with the help of FEDs is indeed possible and seems to lead to plausible results. The results of the fully molecular TGT^{CH3}/amq^{CH3} in the crystal structure and MD relaxed variant are also supporting the conclusions drawn from the united atom variant.

Concluding remarks

In the quest for designing optimal ligands a step in the right direction was shown in this study. This was grounded on an in-depth analysis of the technical and numerical subtleties of such calculations (assessment of grid sizes and closure relations on the RET/AD80 system). After these challenges could be tackled, a workflow was established and directly applied to a model system consisting of the TGT/amq complex, this allowed the revelation of interesting and counter intuitive design directions. On the one hand, this led to several *a priori* design ideas for the concrete derivatisation of the amq-ligand (exchange C3 with N and/or exchange of methyl with ethyl or the introduction of thiol/amine group) that could lead to optimised binding characteristics for the TGT/amq complex system. On the other hand, an explanation for the better binding affinity of the the amq^{H} ligand could be given. The findings of this study should be further backed up by other means, like TI or the 3D RISM-*uv* based method that was used in Ref. [220] and ultimately experimental confirmation is of paramount importance to assess the true potential this methods has.

The field of ligand optimisation is rife with opportunities and 3D RISM-*uu* can play a significant role, if some of the weaknesses described in this work can be addressed. For example: To overcome the systematic shortcomings, overestimation of the desolvation penalty, of the partial *holo* calculations it would be a good idea to not only place the atom at the original ligand position, but to probe its surroundings. This would also solve another inherent problem: placing and calculating FEDs only at the positions of the original atom disregarding the changes in bond lengths upon introduction of another atom or group.

Something which is evident from the presented data, is that FEDs with regard to the different force field parameters can lead to contrasting suggestions for the actual design direction, which in turn leaves room for decisions. It was also shown that the partial *holo* calculations seem to be able to discriminate between the two binders and after accounting for the hysteresis effect the experimental trend could be reproduced.

5 Novel scoring function based on 3D RISM-*uu* and machine learning

5.1 Introduction

In this chapter a new scoring function based on 3D RISM-*uu* and deep neural networks or gradient boosted trees is proposed and evaluated.

The study is designed to show that the addition of atomwise PMFs as an input to a scoring function improves the resulting model with regard to a model that was solely trained on molecular fingerprints. As training data a subset of the PDBbind ("refined set" as defined in Ref. [53]) and as test data, the respective "core[53] set" will be used. The input data for the different models will be either comprised of structural information only, in the form of circular Morgan Fingerprints,[61] calculated by `RDkit`[221] (version 2016.09.4) or the same fingerprints with added atomwise PMF values, calculated by 3D RISM-*uu*.

5.2 Computational details

5.2.1 Structure preparation

The PDBbind[53] 2015 refined set contains 3706 structures, which were stripped of the remaining crystal water. If more than one conformation of the protein was deposited, the most populous conformation was chosen, which is included in the PDB-file. For the parametrisation

the ff14SB[222] force field of the AMBER14[213] package was used for the proteins, GAFF 1.5[208, 209] for the ligands, and parameters of Li et. al.[223] were used for divalent ions. Partial charges of the ligands were calculated using the AM1-BCC method.[210, 211] The same procedure was done for the 2014 core set of the PDBbind, which contains 195 complexes. All structures and parameter files generated can be found in the electronic appendix.

5.2.2 Workflow for RISM-*uv*/*uu* calculations

As a basis for all following RISM calculations, the χ-function (result of 1D RISM-*vv*) was calculated with the dielectrically consistent (DRISM/HNC) theory[190, 191] for pure water (modified TIP3P, see chapter 3). For further details about the generation of the χ-function the reader is referred to page 50. The sole difference between the 1D RISM-*vv* calculations and the 1D RISM-*uv* of the ligand atoms is the maximum residual norm for the DIIS convergence criterion which was set to 10^{-7} for the former calculations and to 10^{-5} for the latter. Throughout all 1D RISM-*uv* calculations the PSE2 closure was used.

For the necessary 3D RISM-*uv* calculations of the proteins, the grid size was automatically chosen to encompass the full complex with a margin of 20 Å in all directions and the grid spacing was set to 0.3 Å. Long range electrostatics were evaluated using the PME[215, 216] of order 8 and short range interactions were cut at 14 Å. Additionally monopole renormalization was used[28, 130] for every calculation. As the convergence criterion the maximum residual norm of the direct correlation functions was set to 10^{-4} and 12 DIIS vectors were used to accelerate the convergence. For all 3D RISM-*uv* calculations, the PSE2 closure relation was used.

All RISM-*uu* calculations were done using the aforementioned 1D/3D RISM-*uv* calculations as a basis and thermodynamic variables were held constant. The atomwise PMFs were calculated

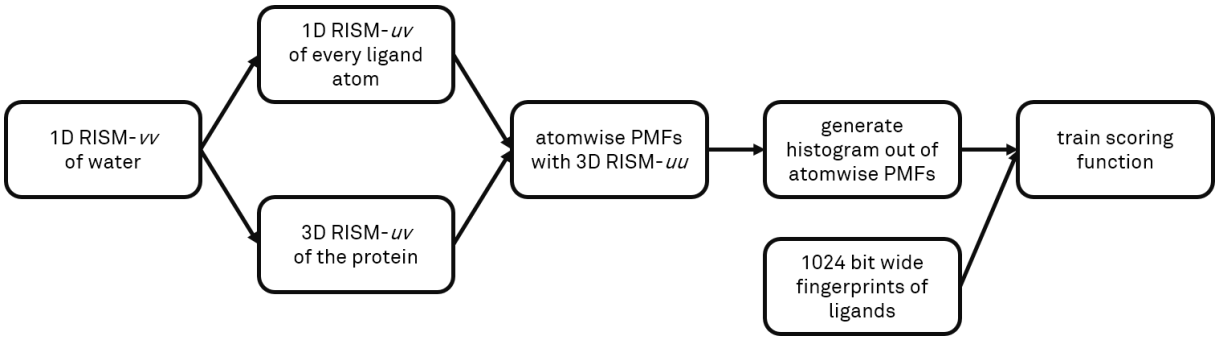

Figure 22: Simplified workflow for the generation of scoring functions based on 3D RISM-*uu*.

according to the so-called *apo* scheme of the previous chapter. To recapitulate: all ligand atoms are placed (at the original coordinated of the *holo* form) in the empty binding site and the PMF is calculated as described in chapter 2.3 The workflow used in this work is shown in Figure 22.

5.2.3 Scoring function generation

Due to problems regarding the automated parametrisation, convergence problems for the 1D RISM-*uv* calculations and technical issues, the final dataset for training and validation purposes consisted of 1321 complexes of the refined set and 54 complexes of the core set. In the refined set, the experimental data was comprised of either K_i (705 complexes) or K_d (616 complexes) values. The core set consisted of 34 K_i values and 20 K_d values. The input of the scoring function generation was either structural ligand information in the form of circular Morgan Fingerprints alone or circular Morgan Fingerprints in conjunction with PMFs calculated by 3D RISM-*uu*.

Now, some closing words about the dataset composition and the distribution of binding affinities: it would be desirable for the used methods if the experimental binding data would be distributed uniformly, which is clearly not the case (see Figure 23). Complexes with really high and really low affinities are underrepresented.

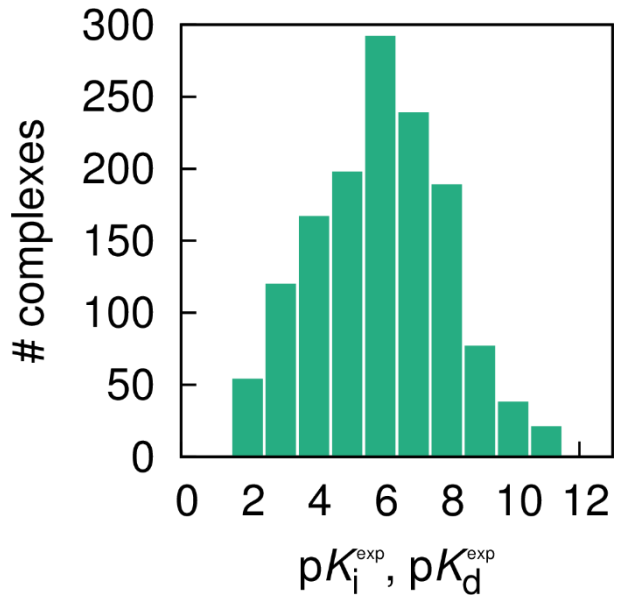

Figure 23: Shows the distribution of binding affinities in the training data set. Complexes with high and low affinities are underrepresented.

Table 5: Hyperparameters of the deep neural networks used for the scoring functions. The parameters are the result of extensive optimisation by the hyperband[224] algorithm.

hyperparameter	DNN1	DNN2
number of layers	3	4
activation function	tanh	tanh
number of hidden units per layer	250	150
weight initialisation	uniform	uniform
L2-regularisation[a]	0.1	0.1
dropout[b]	0.5	0.6
learning rate[c]	0.002	0.002
batch size	60	60
training epochs	2500	2500

[a]additive term that reduces overfitting; [b]Srivastava *et al.*[225] reduces overfitting; [c]step size for the gradient descent

The Morgan Fingerprints were calculated by RDkit[221] (version 2016.09.4) with a radius of 4 and a bit length of 1024. Because the size of the feature vector has to be constant, for the machine learning methods used in this work. The PMFs for the scoring function were represented in a histogram. This was done in the following manner: the overall range of the histogram was chosen to be symmetric from -200 kcal mol^{-1} to +200 kcal mol^{-1} with all values smaller or bigger than that bundled into one bin. Afterwards, 500 linearly spaced bins were created (see Figure 24). The machine learning libraries that were used are `Keras`[226] with the `Theano`[169–171] backend for the neural networks based scoring functions and for the XGBoost[99] models, the library of the same name. All predictions can be found in the electronic appendix for chapter 5.

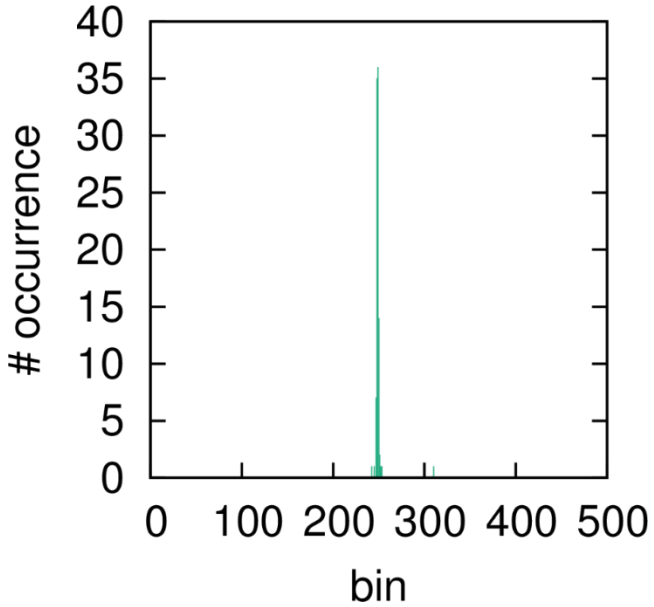

Figure 24: Example for a PMF histogram generated for the scoring functions.

Table 6: Hyperparameters for the XGBoost model used in this work.

hyperparameter	
number of estimators[a]	1000
maximum depth[b]	15
L1-Regularisation	0.01
L2-Regularisation	0.01
γ[c]	0.3
subsample[d]	0.3
minimum child weight[e]	4

[a]number of boosted trees; [b]maximum depth of each tree; [c]minimum loss reduction required for further partitioning of leaf node; [d]fraction of observations that are randomly sampled for each tree; [e]controls overfitting

Hyperparameters for the machine learning methods were chosen after extensive optimisation either with the hyperband[224] algorithm and a random 90/10 split into training/validation set for the neural networks or an exhaustive grid search in conjunction with 5-fold cross-validation for XGBoost. The resulting parameter sets that were used for all further binding affinity predictions are shown in Table 5 for the deep neural networks and in Table 6 for the XGBoost model. For the final model generation the respective training set was randomly split into a training set consisting of 90% of the data and validation set that consisted of 10% of the data, which was used during training as a "early stopping" criterion. "Early stopping" can be understood as a convergence criterion and helps to reduce overfitting. The assignment to the respective set (training or validation) was done randomly. The core set was solely used as a test set and for comparison of the predictive capabilities of the different models. In a last approach, the best three models were combined in a so-called bagging approach by averaging over their predictions.

5.3 Results and discussion

This chapter can be split into three parts: First, the results for the scoring function generated with the deep neural networks are shown. This is followed by the results of the XGBoost model. At last the results of the bagging approach are presented. To further assess the quality of the models, it was calculated if the models are able to distinguish between relatively better binders by calculating:

$$\Delta \mathrm{p}K_x = (\mathrm{p}K_{x,i} - \mathrm{p}K_{x,j}) \forall (i,j) \in \text{complexes}\{\text{exp.}, \text{calc.}\} \tag{77}$$

where $\Delta \mathrm{p}K_x$ is the pairwise differences between the experimental pK values or the computed pK values. By an elementwise comparison of the sign of $\Delta \mathrm{p}K_x$ for the experimental data and calculated data it could be determined how well the scoring functions are able to predict relative changes between two molecules.

Table 7: Performance metrics for the different scoring functions trained with the neural networks, XGBoost or the "bagged" approach on the test dataset (core set PDBbind). The best value for every column is emphasized in bold.

model	dataset subgroup	feature composition	*R*	*p*-value	RMSE	slope	intercept	% trends right
DNN1	whole	fingerprint + PMF	0.66	$4.9 \cdot 10^{-8}$	1.56	0.41	3.59	74.6
DNN1	only K_i	fingerprint + PMF	0.56	$6.5 \cdot 10^{-4}$	1.78	**0.44**	**3.43**	68.6
DNN1	only K_d	fingerprint + PMF	0.44	$5.2 \cdot 10^{-2}$	1.91	0.30	4.11	65.6
DNN1	whole	fingerprint	0.48	$2.1 \cdot 10^{-4}$	1.87	0.33	3.90	66.1
DNN1	only K_i	fingerprint	0.54	$9.9 \cdot 10^{-4}$	1.75	0.37	3.88	66.7
DNN1	only K_d	fingerprint	0.38	$9.5 \cdot 10^{-2}$	1.97	0.25	4.43	64.2
DNN2	whole	fingerprint + PMF	**0.68**	$1.8 \cdot 10^{-8}$	**1.54**	0.41	3.60	74.3
DNN2	half	fingerprint + PMF	0.53	$2.8 \cdot 10^{-5}$	1.75	0.29	4.09	-
DNN2	only K_i	fingerprint + PMF	0.56	$6.1 \cdot 10^{-4}$	1.73	0.39	3.71	69.2
DNN2	only K_d	fingerprint + PMF	0.46	$4.1 \cdot 10^{-2}$	1.83	0.29	4.08	66.3
DNN2	whole	fingerprint	0.52	$6.5 \cdot 10^{-5}$	1.80	0.32	4.16	66.5
DNN2	half	fingerprint	0.49	$1.4 \cdot 10^{-4}$	1.81	0.28	4.38	-
DNN2	only K_i	fingerprint	0.54	$1.0 \cdot 10^{-3}$	1.77	0.38	3.79	66.7
DNN2	only K_d	fingerprint	0.37	$1.1 \cdot 10^{-1}$	1.94	0.22	4.41	61.0
XGBoost	whole	fingerprint + PMF	0.70	$4.2 \cdot 10^{-9}$	1.60	0.30	4.03	75.8
XGBoost	half	fingerprint + PMF	0.66	$3.3 \cdot 10^{-8}$	1.64	0.27	4.22	-
XGBoost	only K_i	fingerprint + PMF	0.75	$2.8 \cdot 10^{-7}$	1.53	**0.33**	3.91	**77.9**
XGBoost	only K_d	fingerprint + PMF	0.71	$4.6 \cdot 10^{-4}$	1.51	**0.33**	**3.52**	75.3
XGBoost	whole	fingerprint	0.66	$5.3 \cdot 10^{-8}$	1.68	0.24	4.35	74.1
XGBoost	half	fingerprint	0.62	$3.6 \cdot 10^{-7}$	1.69	0.24	4.42	-
XGBoost	only K_i	fingerprint	**0.81**	$8.7 \cdot 10^{-9}$	**1.50**	0.32	3.93	77.0
XGBoost	only K_d	fingerprint	0.59	$6.5 \cdot 10^{-3}$	1.69	0.22	4.11	66.8
"bagged"	whole	fingerprint + PMF	**0.71**	$2.5 \cdot 10^{-9}$	**1.52**	0.37	**3.74**	76.1
"bagged"	only K_i	fingerprint + PMF	0.63	$5.7 \cdot 10^{-5}$	1.59	**0.38**	3.68	72.4
"bagged"	only K_d	fingerprint + PMF	0.54	$1.4 \cdot 10^{-2}$	1.70	0.30	3.90	67.9
"bagged"	whole	fingerprint	0.57	$5.7 \cdot 10^{-6}$	1.70	0.30	4.13	69.5
"bagged"	only K_i	fingerprint	0.63	$7.1 \cdot 10^{-5}$	1.61	0.35	3.90	69.9
"bagged"	only K_d	fingerprint	0.45	$4.8 \cdot 10^{-2}$	1.82	0.23	4.32	67.4

In Table 7, the results for the scoring functions trained on the core set with the two different networks are shown. From the data it is evident that the neural network results scale rather

strong with the dataset size. This is expected because it is known that neural networks and in particular deep neural networks can scale to huge datasets of several million entries very well.[227] This was further investigated by cutting the training dataset into halves and retraining one model for both feature sets. For the best model trained on the fingerprint + PMF information, the performance measured by the Pearson correlation coefficient *R* fell from 0.68 to 0.53 and for the model trained on the fingerprint information alone from 0.52 to 0.49. The two best models were trained on the whole dataset, with the fingerprints and added PMF information calculated by 3D RISM-*uu*, and have a *R* of 0.66 for DNN1 and 0.68 for DNN2. The model (DNN2) also has the best RMSE with 1.54. All models have trouble predicting (see Figure 26) really low and high binding affinities, which becomes evident from their respective slopes and intercepts.

Table 8: Performance metrics for the different scoring functions trained with the neural networks, XGBoost or the "bagged" approach on the training dataset (refined set PDBbind). The best value for every column is emphasized in bold.

model	dataset subgroup	feature composition	*R*	*p*-value	RMSE	slope	intercept
DNN1	whole	fingerprint + PMF	0.88	0.0	0.86	0.74	1.69
DNN1	only K_i	fingerprint + PMF	**0.93**	0.0	**0.67**	**0.85**	**1.04**
DNN1	only K_d	fingerprint + PMF	0.91	0.0	0.73	0.78	1.37
DNN1	whole	fingerprint	0.83	0.0	1.02	0.70	1.68
DNN1	only K_i	fingerprint	0.92	0.0	0.71	0.79	1.38
DNN1	only K_d	fingerprint	0.90	0.0	0.79	0.76	1.52
DNN2	whole	fingerprint + PMF	0.88	0.0	0.88	0.69	1.91
DNN2	only K_i	fingerprint + PMF	**0.93**	0.0	0.68	0.81	1.27
DNN2	only K_d	fingerprint + PMF	0.91	0.0	0.76	0.73	1.53
DNN2	whole	fingerprint	0.85	0.0	0.94	0.70	1.90
DNN2	only K_i	fingerprint	0.92	0.0	0.71	0.79	1.36
DNN2	only K_d	fingerprint	0.90	0.0	0.78	0.76	1.42
XGBoost	whole	fingerprint + PMF	**0.90**	0.0	0.96	0.58	2.37
XGBoost	only K_i	fingerprint + PMF	**0.90**	0.0	**0.94**	**0.60**	**2.30**
XGBoost	only K_d	fingerprint + PMF	0.89	0.0	0.95	0.57	2.32
XGBoost	whole	fingerprint	0.87	0.0	1.03	0.54	2.58
XGBoost	only K_i	fingerprint	0.87	0.0	1.02	0.56	2.59
XGBoost	only K_d	fingerprint	0.86	0.0	1.04	0.52	2.60
"bagged"	whole	fingerprint + PMF	0.90	0.0	0.85	0.67	1.99
"bagged"	only K_i	fingerprint + PMF	**0.93**	0.0	**0.69**	**0.75**	**1.54**
"bagged"	only K_d	fingerprint + PMF	0.92	0.0	0.76	0.70	1.74
"bagged"	whole	fingerprint	0.88	0.0	0.90	0.65	2.05
"bagged"	only K_i	fingerprint	0.92	0.0	0.76	0.71	1.78
"bagged"	only K_d	fingerprint	0.90	0.0	0.80	0.68	1.85

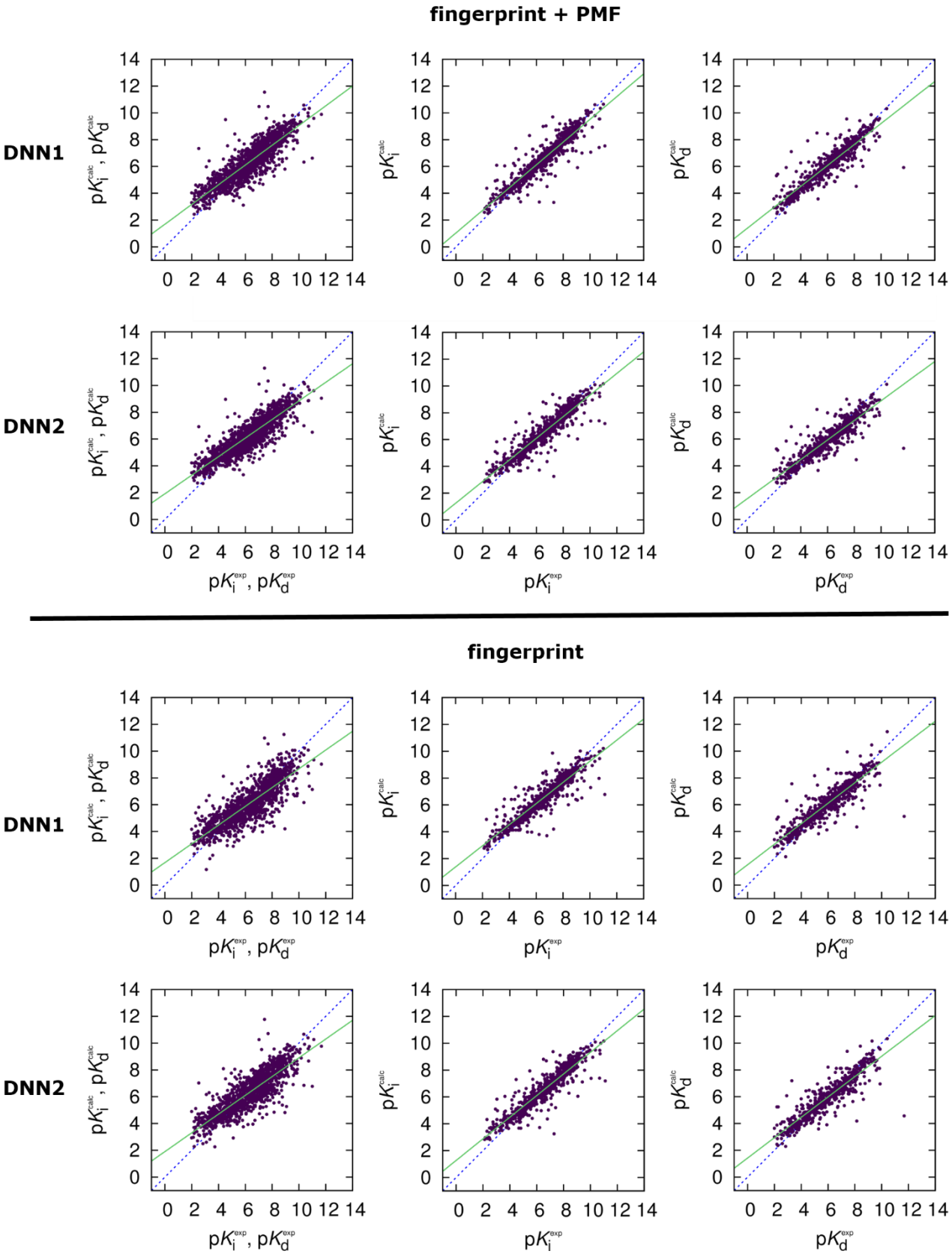

Figure 25: Calculated and experimental data for the training set (refined set) plotted against each other for the neural network models trained in this work. The top row and second row is comprised of models that were trained on fingerprint + PMF information, the bottom rows show models that were trained only on the fingerprint data. From left to right the models were trained on the whole dataset, K_i dataset or K_d dataset.

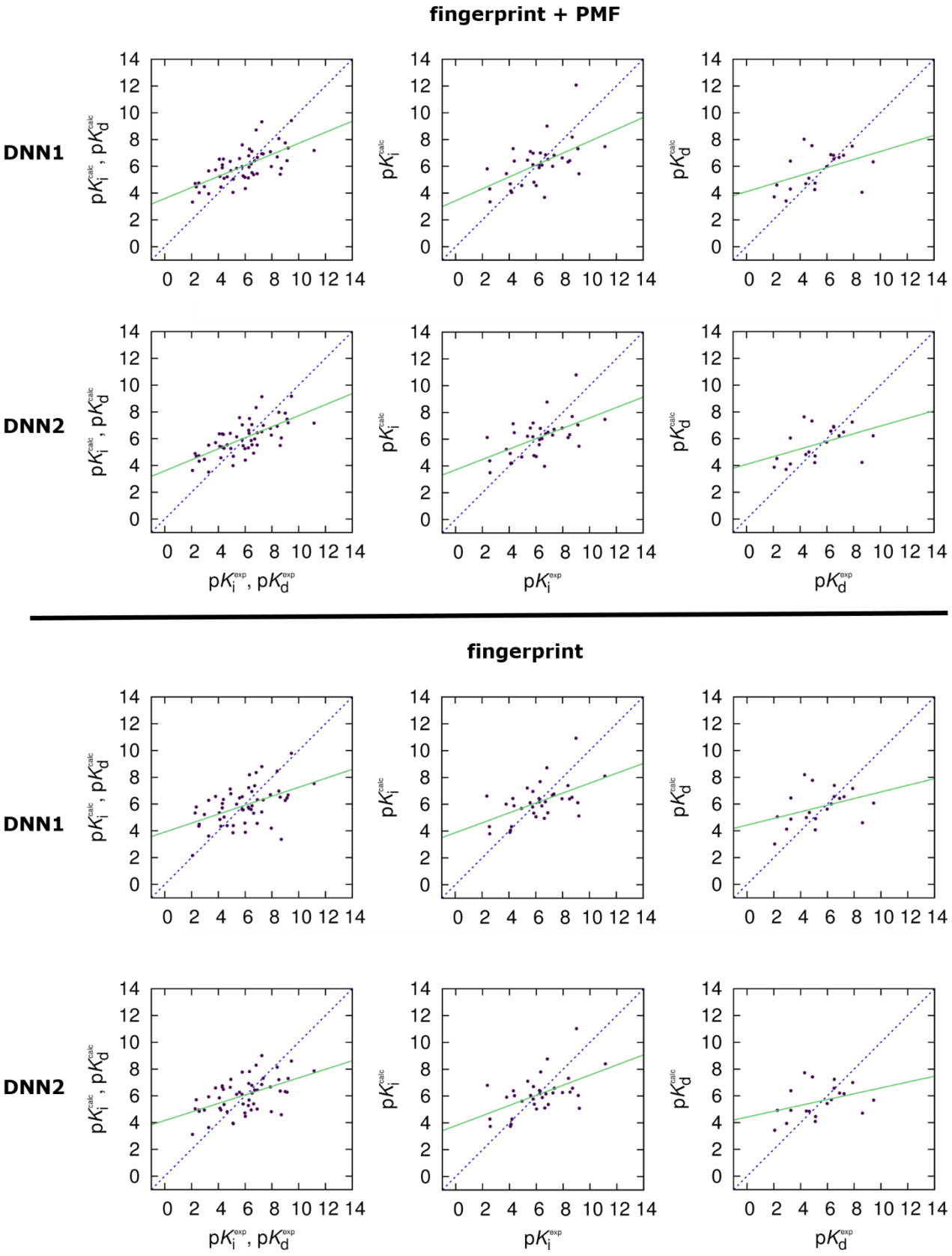

Figure 26: Calculated and experimental data for the test set (core set) plotted against each other for the neural network models trained in this work. The top row and second row is comprised of models that were trained on fingerprint + PMF information, the bottom row shows models that were trained only on the fingerprint data. From left to right the models were trained on the whole dataset, K_i dataset or K_d dataset.

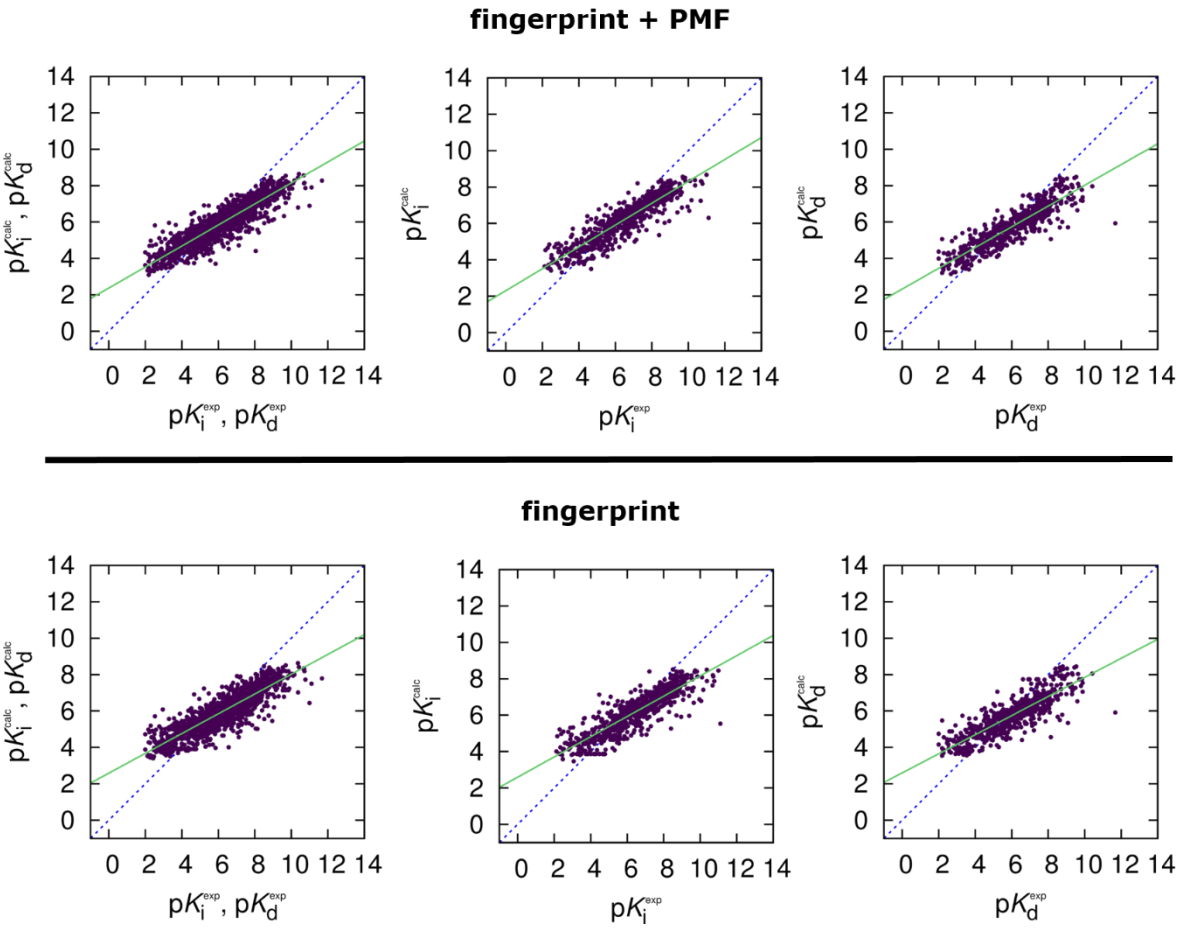

Figure 27: Calculated and experimental data for the training data (refined set) plotted against each other for the XGBoost models trained. The top row shows models that were trained on fingerprint + PMF information, the bottom row shows models that were trained only on the fingerprint data. From left to right the models were trained on the whole dataset, K_i dataset or K_d dataset.

The deep neural network models really benefit from the added PMF data. The best model that was trained solely on the fingerprint information reaches an *R* of 0.54 and an RMSE of 1.77 (core set). From the models trained on the pure datasets, that only incorporated either K_i or K_d data, the models trained on the former show the better performance. The models trained with the added PMF data slightly outperform the models trained only on the fingerprint data. The worst performing models were trained on the K_d data only, in particular the model trained on the fingerprint information alone that only reached an *R* of 0.37 and RMSE of 1.94. One possible explanation lies within the slightly bigger dataset for the K_i data (705 datapoints for K_i and 616 datapoints for K_d). It is interesting that the best DNN models are able to predict the low and high affinity complexes rather well, although these complexes are underrepresented in the training set (see Figure 23). This hints probably at an underlying generalisation capability, which means a good transferability of the model to unseen data that would be much desired. All models are able to predict 61.0 % to 74.6 % of the relative trend changes within their respective datasets (core set). The best models in this respect are DNN1 and DNN2 trained on

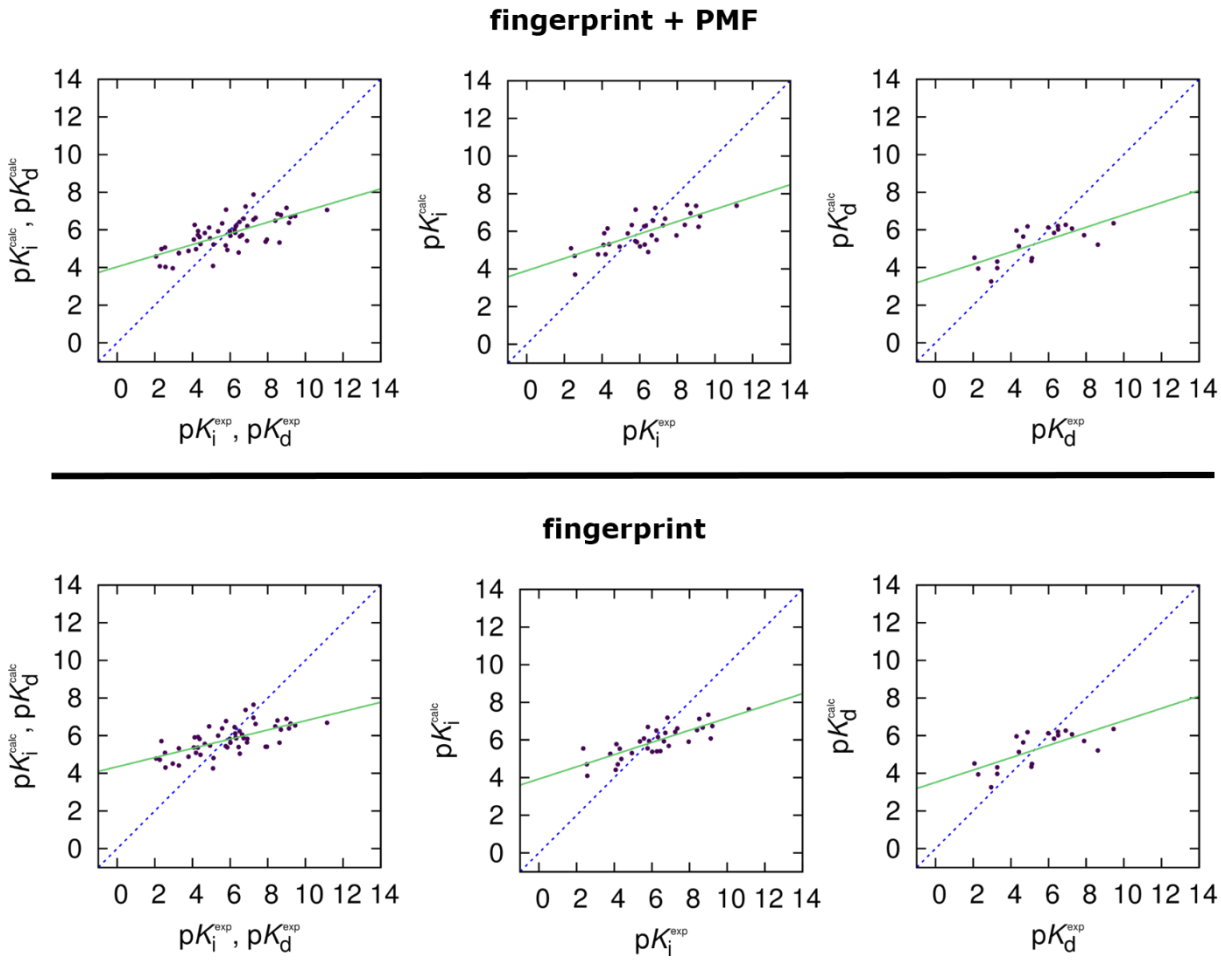

Figure 28: Calculated and experimental data for the test data (core set) plotted against each other for the XGBoost models trained. The top row shows models that were trained on fingerprint + PMF information, the bottom row shows models that were trained only on the fingerprint data. From left to right the models were trained on the whole dataset, K_i dataset or K_d dataset.

the whole dataset and fingerprint + PMF information which can predict 74.6 % or 74.3 % of the correct trends, respectively.

Next a look at the outliers is warranted, for which the corresponding prediction data is shown in Table 13, Table 14, and Table 15 of the appendix. For the whole data set the DNN models overestimate the binding affinity of the complexes with the pdbcode 1UTO and 3G2Z the most, with the experimental pK_i being 2.27 and 2.36 respectively. The calculated pK_i lie in the range of 4.73 – 4.89 for 1UTO and 5.71 – 5.78 for 3G2Z. Both complexes have small ligands in the range 122 g/mol to 179 g/mol. The proteins that form the aforementioned complexes both belong to the hydrolase family. The complexes where the binding affinity is underestimated the most are 1HFS and 1MQ6, where the experimental pK_i is 8.70 and 11.15 respectively. The calculated pK_i range between 3.37 – 4.59 for 1HFS and 7.16 – 7.19 for 1MQ6. Both ligands have higher molecular weights between 586 g/mol and 705 g/mol. The proteins that form the complexes are a hydrolase in the case of 1HFS and Factor Xa in the case of 1MQ6.

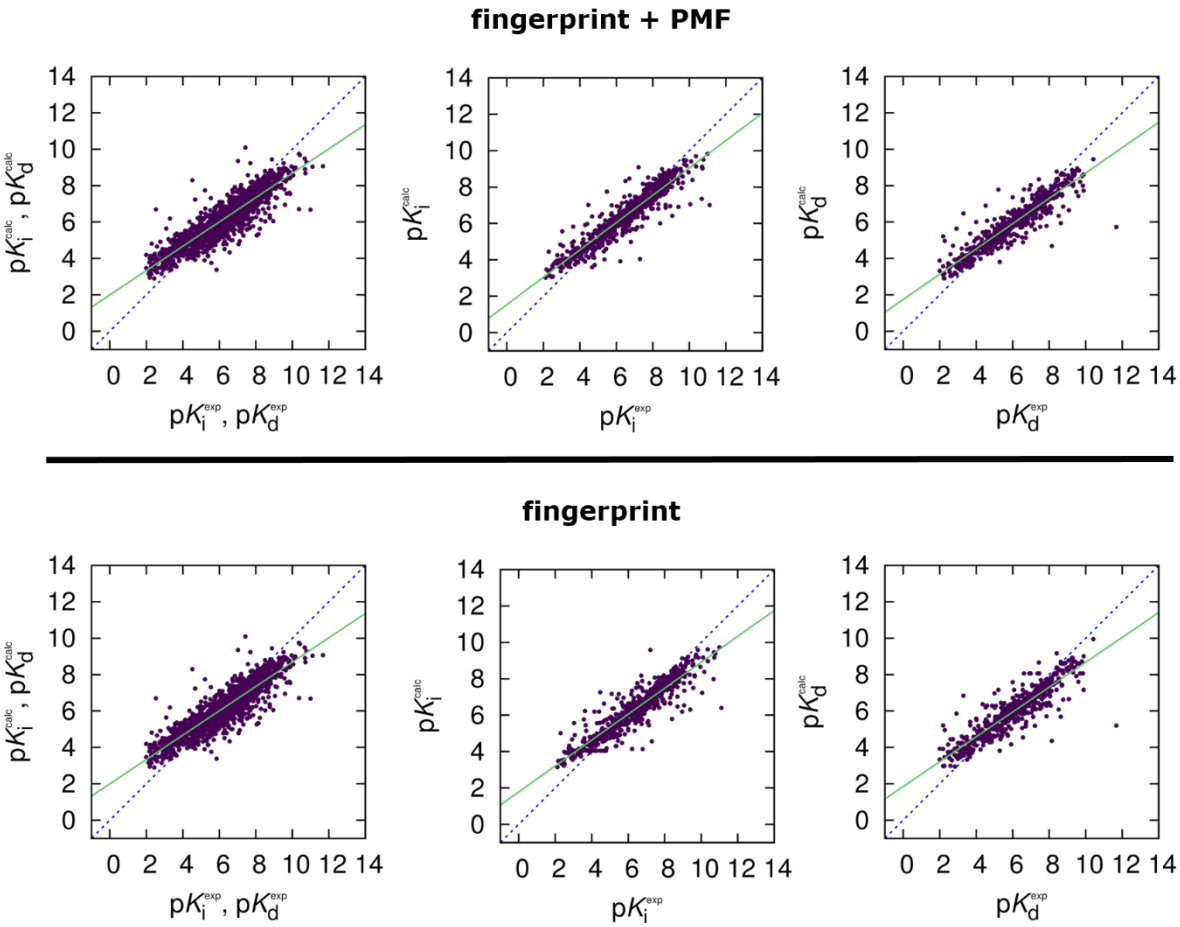

Figure 29: Calculated and experimental data for the training set (refined set) plotted against each other for the "bagged" models. The top row results from the bagging of the models that were trained on fingerprint + PMF information, the bottom row shows models that were bagged on models that were trained on fingerprint data alone. From left to right the models are shown for the whole dataset, K_i dataset or K_d dataset.

One common denominator between all these ligands is that they all have more or less extended π-systems. This could point into the direction of a systematic problem because the ansatz chosen for this work relies heavily on atomwise PMFs calculated in the *apo* binding site, where every information about aromaticity is lost. Thus the only information the models have about the aromaticity of the ligands is provided by the fingerprints which could be not sufficient for a more accurate prediction.

To investigate the possibility of overfitting the models were also applied to their respective training data set (refined set). All DNN models perform significantly better, with regard to all performance metrics, on their respective training data set. Interestingly the best performance is achieved on the K_i dataset with the added PMF data, where DNN1 and DNN2 reach an R of 0.93 and RMSEs of 0.67 and 0.68 respectively. Is this a sign that the underlying experimental data is more reliable for the K_i data? The better performance on the training data is also a hint that overfitting could be a problem for the DNN models.

Now coming to the results for the trained XGBoost model, which are summarized in Table 7 and Figure 28. The XGBoost model also seems to benefit from a larger dataset, which is seen by comparison of the performance metrics for the whole and the half dataset with their respective feature sets. In detail, this means that the *R* of the model that was trained on half of the dataset and the fingerprint + PMF data falls from 0.70 to 0.66 and the RMSE rises from 1.60 to 1.64. If the model is trained only on the fingerprint data the *R* falls from 0.66 to 0.62 and the RMSE rises slightly from 1.68 to 1.69 (on core set). Interestingly, the overall best result for the trained XGBoost models was achieved on the dataset containing only the K_i data and fingerprints: this model reaches an *R* of 0.81, an RMSE of 1.50 and predicts 77.0 % of the correct trends (core set).

The best model that was trained on the fingerprint + PMF data was also trained on the K_i data alone. It has an *R* of 0.75, an RMSE of 1.53 and predicts 77.9 % of the trends. The worst

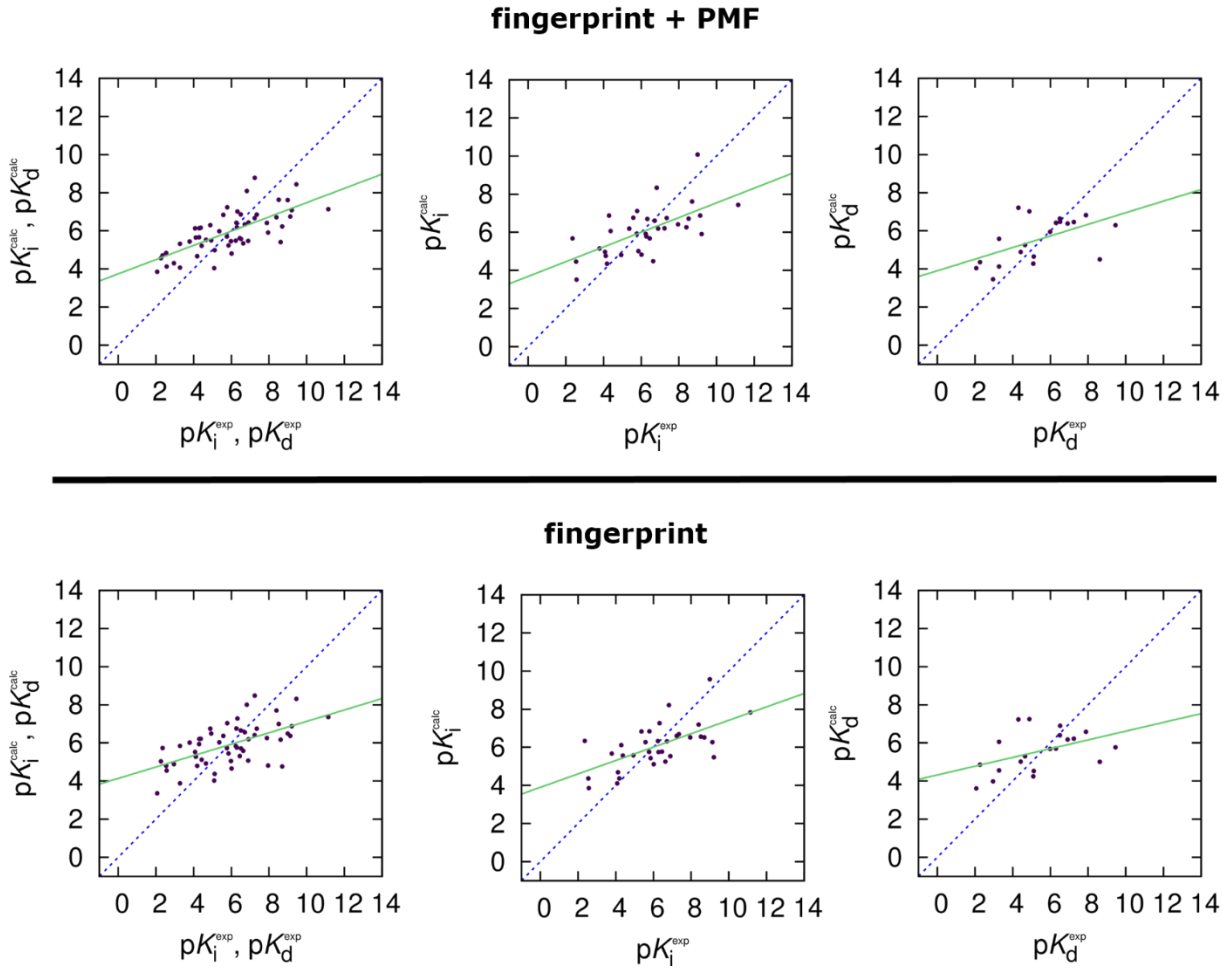

Figure 30: Calculated and experimental data for the test set (core set) plotted against each other for the "bagged" models. The top row results from the bagging of the models that were trained on fingerprint + PMF information, the bottom row shows models that were bagged on models that were trained on fingerprint data alone. From left to right the models are shown for the whole dataset, K_i dataset or K_d dataset.

models trained on both feature sets were trained on the K_d data alone and achieve for the fingerprint feature set an *R* of 0.59, an RMSE of 1.69 and predict 66.8 %. For the "fingerprint + PMF" feature set, the worst model has an *R* of 0.71, an RMSE of 1.51 and predicts 75.3 % of the correct trends. Apart from the K_i, dataset the performance metrics for the models trained on the whole and on the K_d data benefit from the added PMF data. For example, for the whole dataset added PMF data leads to a better *R* by 0.04, a better RMSE by 0.08 and a change in the slope and intercept that are also favourable. For the K_d dataset these trends are even bigger.

Comparing the XGBoost models with their respective DNN models, shows that the XGBoost models produce a tighter distribution with a higher correlation coefficient. For the RMSEs the picture changes only in respect to the best DNN (DNN2/whole dataset/ fingerprint + PMF) model which achieves a similar or even better RMSE than the XGBoost models. A look at the slopes of the linear regression reveals that the DNN models seem to perform better with slopes that are slightly better than the XGBoost models. The consequence of the weak slopes of the XGBoost models is that they are not very sensitive in regard to really strong and weak binder, but are able to discriminate between binders and non-binders rather well.

Regarding the outliers the same trend is seen for the DNN models holds true for the XGBoost models. They too have problems with molecules that have extended π-systems which could be a problem with the representation of the PMFs in an atomwise manner.

In Table 8 and Figure 27, the data for the XGBoost models on their respective training data set is shown. The XGBoost models also show a higher performance on the training set than on the test, but overall the differences are not as high as for the DNN models. Interestingly, the slopes of the XGBoost models on the training data do not benefit as much as the DNN models, which is in accordance to the performance on the test data were the XGBoost models showed problematic behaviour regarding the slopes. Thus, this could be an inherent weakness of the XGBoost model regarding the non-uniform binding affinity distribution. Overall the XGBoost model seems to be unlikely to suffer from significant overfitting.

To enhance the predictive capabilities of the trained models they were combined in a "bagging" approach similar to Ashtawy *et al.*[54] This was done by computing the arithmetic mean of the predictions for both DNN and the XGBoost models. The results for this are shown in Table 7 and Figure 30.

The dataset and feature combination that benefits most from the bagging approach is the whole dataset in combination with fingerprint + PMF information where an *R* of 0.71, RMSE of 1.52, and a slope of 0.37 is achieved, which is better compared to all other models trained on this dataset. Compared to the DNN models, the "bagged" models show only slightly better results for the *R* and RMSEs. The slopes and intercepts are generally better for the DNN models alone, except for the model that was trained on the whole dataset and fingerprint information here the bagged approach shows the slightly better slope. In case of the XGBoost models, the result of the comparison is mostly inverted: apart from the model that was trained on the whole dataset and fingerprint + PMF information all other XGBoost models show better *R* values than their "bagged" counterparts. The slopes and intercepts of the XGBoost models trained on the K_i data and fingerprint +PMF information are better than the respective bagged model, which is also the case for the intercept of the model that was solely trained on the fingerprint and the K_d data.

Now, the question arises how the presented results compare to other work that was done in this field. In Table 9, a comparison between other machine learning based scoring functions, classical scoring functions and the models presented in this work is shown. The best performing models are either based on neural networks, combined through a "boosting" (BsN-Score) or "bagging" (BgN-Score) approach or based on a random forest (RF). As feature set all these models used a "meta" set comprised of descriptors taken from X-Score[38] (X), AffiScore[228] (A), Gold[36] (G), and RF-Score[45] (R).

Table 9: Comparison of other scoring functions to the best models trained in this work. All scoring functions were tested on the core set or a subset of the PDBbind. The models calculated in this work are underlined.

Machine learning based	***R***	**RMSE**
BsN-Score::XARG	0.82[a]	1.38
BgN-Score::XARG	0.80[a]	1.45
RF::XARG	0.79[a]	1.50
Bag[c]	0.71	1.52
XGBoost[c]	0.70	1.60
DNN2[c]	0.68	1.54
SNN-Score::X	0.68[a]	1.76
Classical		
X-Score::HMScore	0.64[a]/0.61[b]	1.87
DrugScoreCSD	0.57[a]/0.53[b]	-
SYBL::ChemScore	0.55[a]/0.59[b]	-

[a]Ashtawy et. al.[54], [b]Li et. al.[13] test set based on PDBbind core set, [c]this work

The models presented in this work are competitive with the top performing scoring functions and outperform classical scoring functions significantly. It has to be noted that the models in this work could only be applied to a subset of the core set of the PDBbind (see computational details). The benefit of the features chosen in this work is that, in case of the PMF data, they retain a degree of chemical and physical interpretability. For the top performing machine learning models (BsN-Score, BgN-Score, RF) this is not the case, because they use such a wide set of (overlapping) features. Thus the ansatz chosen in this work seems to be promising and leaves much room for improvement.

5.4 Concluding remarks

This chapter showed that it is not only possible to derive a scoring function, based on fingerprints and PMFs calculated by 3D RISM-*uu*, but that the models presented in this work are competitive with other scoring functions, that are based on machine learning techniques and even outperform classical scoring functions. This conclusion was gained after training several models with varying compositions of the underlying dataset and two different feature sets that either contained only structural information, based on fingerprints or fingerprints + PMF information calculated by 3D RISM-*uu*. The training dataset itself was composed of a subset of the PDBbind refined set (1321 complexes). The external test set was a subset of the corresponding core set (54 complexes). The scoring functions were either trained with deep neural networks or with a boosted regression tree method, called XGBoost. The best result that was achieved with the whole dataset was gained by a bagging approach, which is simply the arithmetic mean of all models trained and reached a Pearson correlation coefficient of 0.71 and an RMSE of 1.52.

Nonetheless, there is a lot of work to be done. First, the distribution of binding affinities is far from uniform, which would be desirable for the methods being used in this work. A more uniform distribution could be enforced by taking all complexes with very low/high binding affinities and taking an appropriate amount of complexes with medium binding affinities in a way that the subset of the data resembles a uniform distribution. Then this process needs to be repeated until all complexes with medium binding affinities are allotted to a subset. This would generate several subsets which could then be used for further training purposes. The second

problem is the overall size of the dataset. It would be really beneficial if more data was available, which would probably be the single most important factor for the generation of better models.

On the architectural front of the machine learning methods employed several ideas come to mind: For example, convolutional neural networks could be used for the prediction, which are generally able to extract features of interest and were already applied in the field of binding affinity predictions (see Ref . [57]).

On the physics side of the equation, the 3D RISM-*uu* framework needs to be advanced to the fully molecular equation, which would then allow oneself to feed the calculated molecular PMF information, which is already on a grid directly into a convolutional deep neural network. Further, the results indicate that aromaticity is a problem for the models trained in this work. This could be due to the atomwise PMFs that are based on calculations in the *apo* binding site. It would thus be possible to use the partial *holo* that was presented in chapter 4, which would be able to account for effects based on the aromaticity of the ligand. This would lead to a significantly higher computational demand in respect to the *apo* calculations, but would still be faster than TI-MD calculations. The scoring function generation could also be coupled with the FED information, which could possibly lead to better results.

With the ability to distinguish between the quality of two binders in 76 % of the cases, one could envision a *in silico* automated molecule optimizer that couples a decent scoring function with an algorithm, which optimizes the binding affinity. One possible optimisation algorithm could be a "reinforcement learner." Reinforcement learning recently achieved fame with the Go playing program AlphaGo[229] by DeepMind, which was the first program to defeat a top human Go player. AlphaGo relied heavily on reinforcement learning and in particular on "Deep Q-learning" and variants thereof.[230–233] In simple terms, a reinforcement learning algorithm tries to find a policy that maximizes a future reward signal. This is done in the following way: The algorithm starts by taking a more or less random action. For the *in silico* automated molecule optimizer this would mean to change an atom or group in the molecule to optimize. This change is evaluated by the "environment." Here comes the scoring function into play, which ranks both molecules against each other. If the change leads to a better binder, relatively to the old one, the algorithm gets a reward which gets fed back into a deep neural network, which tries to maximize future rewards. This is done iteratively and the random steps are downregulated, but never completely (balance between "exploration" and "exploitation") replaced by the learned policy

that hopefully maximizes rewards and therefore maximizes the binding affinity of a molecule to its target.

Another possible route would be to use "generative adversarial networks" (GANs) which were invented by Goodfellow[234]. They are gaining popularity[235, 236] within the machine learning community and are often used to create realistic looking images.[237] The principle behind GANs is that two networks compete with each other. One network, the generator, tries to deceive the other network, in this case a scoring function, to misclassify his generated molecules as good binders, thereby generating new and hopefully better binders. A further advantage could be that the model potentially generates molecules which are easy to synthesise because they are heavily influenced by the molecules that are already known.

6 Summary and Conclusion

The aims of this work were to develop and implement novel tools to predict and use PMF data calculated by 3D RISM-*uu*. First, free energy derivatives (FED) were defined within the 3D RISM-*uu* framework and then applied to a model system comprised of the 18-crown-6-ether and a potassium ion. The results from this study were then translated and refined for the usage with two protein ligand systems, namely RET/AD80 and TGT/amq. In the last part of this work a novel scoring function, based on structural fingerprint data and PMFs calculated by 3D RISM-*uu*, was implemented and evaluated.

It was shown that PMF and FED data calculated for the crown ether model system were in reasonable agreement with already published experimental and theoretical data. It was also shown that 3D RISM-*uu* is able to compute the correct topography of the PMF hypersurface compared to explicit TI-MD simulations. A big advantage of the 3D RISM-*uu* ansatz is on one hand the orders of magnitude greater computational efficiency compared to MD based methods and on the other hand the better description of the solvent distribution compared to typical implicit solvent models. A typical calculation with the *apo* scheme takes one to two hours (including the *uv* calculation), one to two days in the partial *holo* scheme and a typical TI-MD would take one to two weeks.

These promising results were then translated to two protein ligand systems (RET/AD80 and TGT/amq). Where it could be shown that calculations with PSE order 2 and a grid spacing of 0.3 Å seem to give the best compromise between the convergence behaviour of the 3D RISM-*uv* calculations and correct (within the method) description of the PMF data. For the TGT/amq system it could be shown that the difference in binding affinity between the amq^{H} and amq^{CU} ligand can be best explained by the change of the partial charges within the molecule (this effect was strongest for the neighbouring atoms). The strong effect of the changed partial charges could mean that the polarizability plays a crucial role in this system. Even after accounting for the hysteresis effect the experimental trend could be reproduced using FEDs, calculated by 3D RISM-*uu*, and a linear approximation scheme. Furthermore, a new design

direction, based on the FEDs calculated with respect to the typical non-bonded force field parameters, could be proposed. This new design direction could possibly lead to an even better binder, which has to be verified by other theoretical methods (e.g. TI-MD) and ultimately by experimental verification.

In the last part of this work a novel type of scoring function based on structural data, in the form of molecular fingerprints, and PMF data calculated by 3D RISM-*uu* was trained using modern machine learning methods. The trained scoring functions showed comparable performance in comparison to other machine learning based scoring function and outperformed classical scoring functions.

In the immediate future the proposed models for the computation of the FEDs should be enhanced by alleviating the problem of the fixed ligand geometry in a practical and computationally feasible way. One idea would be to probe the surrounding of the original ligand atom position and generate the PMF value by integrating over this small volume.

To advance the proposed scoring functions further it would be desirable to have access to the fully molecular 3D RISM-*uu* PMF, which could then be used directly with a convolutional neural network. It would also be desirable to use PMF values calculated with the partial *holo*, which seem to be superior to the *apo* scheme.

For the future one could envision a "virtual molecule design machine" that would be able to predict novel binders. This "virtual designer" would be based on a closed loop: FEDs are calculated for a given molecule and lead to a design direction that is translated into the appropriate chemistry. This new molecule is than scored and gets fed back into the cycle.

7 References

1. J. Woodcock, R. Woosley, The FDA Critical Path Initiative and Its Influence on New Drug Development*. *Annu. Rev. Med.* **59**, 1–12 (2008).
2. J. Polanski, J. Bogocz, A. Tkocz, The analysis of the market success of FDA approvals by probing top 100 bestselling drugs. *J. Comput. Aided. Mol. Des.* **30**, 381–389 (2016).
3. M. S. Kinch, D. Hoyer, A history of drug development in four acts. *Drug Discov. Today*. **20**, 1163–1168 (2015).
4. A. Ganesan, M. L. Coote, K. Barakat, Molecular dynamics-driven drug discovery: leaping forward with confidence. *Drug Discov. Today*. **0** (2016), doi:10.1016/j.drudis.2016.11.001.
5. E. Kim, B. S. Moore, Y. J. Yoon, Reinvigorating natural product combinatorial biosynthesis with synthetic biology. *Nat. Chem. Biol.* **11**, 649–659 (2015).
6. V. Lounnas *et al.*, Current Progress in Structure-Based Rational Drug Design Marks a New Mindset in Drug Discovery. *Comput. Struct. Biotechnol. J.* **5**, 1–14 (2013).
7. L. Ferreira, R. dos Santos, G. Oliva, A. Andricopulo, Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*. **20**, 13384–13421 (2015).
8. S. Grinter, X. Zou, Challenges, Applications, and Recent Advances of Protein-Ligand Docking in Structure-Based Drug Design. *Molecules*. **19**, 10150–10176 (2014).
9. R. D. Taylor, P. J. Jewsbury, J. W. Essex, A review of protein-smallmolecule docking methods 151. *J. Comput. Aided. Mol. Des.* **16**, 151–166 (2002).
10. I. A. Guedes, C. S. de Magalhães, L. E. Dardenne, Receptor–ligand molecular docking. *Biophys. Rev.* **6**, 75–87 (2014).
11. E. Yuriev, M. Agostino, P. A. Ramsland, Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **24**, 149–164 (2011).
12. E. Yuriev, J. Holien, P. A. Ramsland, Improvements, trends, and new ideas in molecular docking: 2012-2013 in review. *J. Mol. Recognit.* **28**, 581–604 (2015).
13. Y. Li, L. Han, Z. Liu, R. Wang, Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *J. Chem. Inf. Model.* **54**, 1717–1736 (2014).
14. A. Biela, M. Betz, A. Heine, G. Klebe, Water makes the difference: rearrangement of water solvation layer triggers non-additivity of functional group contributions in protein-ligand binding. *ChemMedChem*. **7**, 1423–34 (2012).
15. G. Lemmon, J. Meiler, Towards ligand docking including explicit interface water molecules. *PLoS One*. **8**, e67536 (2013).

16. B. Breiten *et al.*, Water Networks Contribute to Enthalpy / Entropy Compensation in Protein-Ligand Binding. *J. Am. Chem. Soc.* **135**, 15579–15584 (2013).

17. R. Baron, P. Setny, J. Andrew McCammon, Water in Cavity−Ligand Recognition. *J. Am. Chem. Soc.* **132**, 12091–12097 (2010).

18. J. E. Ladbury, Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.* **3**, 973–980 (1996).

19. M. S. Bodnarchuk, Water, water, everywhere… It's time to stop and think. *Drug Discov. Today*. **21**, 1139–1146 (2016).

20. *SZMAP* (OpenEye Scientific Software, Inc., Santa Fe, NM, USA, 2015; www.eyesopen.com).

21. *WaterFLAP* (Molecular Discovery, Hertfordshire, UK, 2016; www.moldiscovery.com).

22. A. Amadasi *et al.*, Robust Classification of "Relevant" Water Molecules in Putative Protein Binding Sites. *J. Med. Chem.* **51**, 1063–1067 (2008).

23. T. Young, R. Abel, B. Kim, B. J. Berne, R. A. Friesner, Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proc. Natl. Acad. Sci.* **104**, 808–813 (2007).

24. R. Abel, T. Young, R. Farid, B. J. Berne, R. A. Friesner, Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **130**, 2817–2831 (2008).

25. F. Hirata, P. J. Rossky, B. M. Pettitt, The interionic potential of mean force in a molecular polar solvent from an extended RISM equation. *J. Chem. Phys.* **78**, 4133 (1983).

26. B. M. Pettitt, M. Karplus, The potential of mean force between polyatomic molecules in polar molecular solvents. *J. Chem. Phys.* **83**, 781 (1985).

27. A. Kovalenko, F. Hirata, Potential of Mean Force between Two Molecular Ions in a Polar Molecular Solvent: A Study by the Three-Dimensional Reference Interaction Site Model. *J. Phys. Chem. B.* **103**, 7942–7957 (1999).

28. T. Kloss, S. M. Kast, Treatment of charged solutes in three-dimensional integral equation theory. *J. Chem. Phys.* **128**, 134505 (2008).

29. J. Liu, R. Wang, Classification of current scoring functions. *J. Chem. Inf. Model.* **55**, 475–482 (2015).

30. S. Huang, S. Z. Grinter, X. Zou, Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.*, 12899–12908 (2010).

31. I. Klapper, R. Hagstrom, R. Fine, K. Sharp, B. Honig, Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: Effects of ionic strength and amino-acid modification. *Proteins Struct. Funct. Genet.* **1**, 47–59 (1986).

32. W. Im, D. Beglov, B. Roux, Continuum solvation model: Computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. *Comput. Phys. Commun.* **111**, 59–75 (1998).

33. W. C. Still, A. Tempczyk, R. C. Hawley, T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129 (1990).

34. R. Huey, G. M. Morris, A. J. Olson, D. S. Goodsell, A semiempirical free energy force

field with charge-based desolvation. *J. Comput. Chem.* **28**, 1145–1152 (2007).

35. G. M. Morris *et al.*, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19**, 1639–1662 (1998).

36. G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727–748 (1997).

37. Q. U. Ain, A. Aleksandrova, F. D. Roessler, P. J. Ballester, Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**, 405–424 (2015).

38. R. Wang, L. Lai, S. Wang, Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided. Mol. Des.* **16**, 11–26 (2002).

39. M. D. Eldridge, C. W. Murray, T. R. Auton, G. V Paolini, R. P. Mee, Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided. Mol. Des.* **11**, 425–445 (1997).

40. H. Gohlke, M. Hendlich, G. Klebe, Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **295**, 337–356 (2000).

41. S.-Y. Huang, X. Zou, An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **27**, 1866–1875 (2006).

42. S.-Y. Huang, X. Zou, An iterative knowledge-based scoring function to predict protein–ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **27**, 1876–1882 (2006).

43. J. D. Durrant, J. A. McCammon, NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein−Ligand Complexes. *J. Chem. Inf. Model.* **50**, 1865–1871 (2010).

44. J. D. Durrant, J. A. McCammon, NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *J. Chem. Inf. Model.* **51**, 2897–2903 (2011).

45. P. J. Ballester, J. B. O. Mitchell, A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics.* **26**, 1169–1175 (2010).

46. P. J. Ballester, A. Schreyer, T. L. Blundell, Does a More Precise Chemical Description of Protein–Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* **54**, 944–955 (2014).

47. D. Zilian, C. A. Sotriffer, SFCscore RF : A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **53**, 1923–1933 (2013).

48. H. Li, K.-S. Leung, M.-H. Wong, P. J. Ballester, Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study. *BMC Bioinformatics.* **15**, 291 (2014).

49. H. Li, K.-S. Leung, M.-H. Wong, P. J. Ballester, Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inform.* **34**, 115–126 (2015).

50. O. Trott, A. J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*

31, 455–461 (2009).

51. J. D. Durrant, J. A. McCammon, BINANA: A novel algorithm for ligand-binding characterization. *J. Mol. Graph. Model.* **29**, 888–893 (2011).

52. L. Hu, M. L. Benson, R. D. Smith, M. G. Lerner, H. A. Carlson, Binding MOAD (Mother Of All Databases). *Proteins Struct. Funct. Bioinforma.* **60**, 333–340 (2005).

53. Z. Liu *et al.*, PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*. **31**, 405–412 (2015).

54. H. M. Ashtawy, N. R. Mahapatra, BgN-Score and BsN-Score: Bagging and boosting based ensemble neural networks scoring functions for accurate binding affinity prediction of protein-ligand complexes. *BMC Bioinformatics*. **16**, 12 (2015).

55. J. H. Friedman, Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).

56. D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, L.-X. Zhang, H.-D. Li, The boosting: A new idea of building models. *Chemom. Intell. Lab. Syst.* **100**, 1–11 (2010).

57. I. Wallach, M. Dzamba, A. Heifets, AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *arXiv:1510.02855v1 [cs.LG]* (2015).

58. M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).

59. A. P. Bento *et al.*, The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42**, D1083–D1090 (2014).

60. T. Unterthiner *et al.*, Deep Learning as an Opportunity in Virtual Screening. *Deep Learn. Represent. Learn. Work. NIPS 2014*, 1–9 (2014).

61. D. Rogers, M. Hahn, Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

62. G. Dahl, N. Jaitly, R. Salakhutdinov, Multi-task Neural Networks for QSAR Predictions. *arXiv Prepr. arXiv1406.1231*, 1–21 (2014).

63. J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, V. Svetnik, Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **55**, 263–274 (2015).

64. S. Genheden, T. Luchko, S. Gusarov, A. Kovalenko, U. Ryde, An MM/3D-RISM Approach for Ligand Binding Affinities. *J. Phys. Chem. B.* **114**, 8505–8516 (2010).

65. N. Blinov, L. Dorosh, D. Wishart, A. Kovalenko, Association Thermodynamics and Conformational Stability of β-Sheet Amyloid β(17-42) Oligomers: Effects of E22Q (Dutch) Mutation and Charge Neutralization. *Biophys. J.* **98**, 282–296 (2010).

66. T. Imai, K. Oda, A. Kovalenko, F. Hirata, A. Kidera, Ligand Mapping on Protein Surfaces by the 3D-RISM Theory: Toward Computational Fragment-Based Drug Design. *J. Am. Chem. Soc.* **131**, 12430–12440 (2009).

67. D. Nikolić, N. Blinov, D. Wishart, A. Kovalenko, 3D-RISM-Dock: A New Fragment-Based Drug Design Protocol. *J. Chem. Theory Comput.* **8**, 3356–3372 (2012).

68. Y. Kiyota, N. Yoshida, F. Hirata, A New Approach for Investigating the Molecular Recognition of Protein: Toward Structure-Based Drug Design Based on the 3D-RISM Theory. *J. Chem. Theory Comput.* **7**, 3803–3815 (2011).

69. D. Nikolić, N. Blinov, D. Wishart, A. Kovalenko, 3D-RISM-Dock: A New Fragment-

Based Drug Design Protocol. *J. Chem. Theory Comput.* **8**, 3356–3372 (2012).

70. C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).

71. C. A. Lipinski, Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **1**, 337–341 (2004).

72. F. Mrugalla, S. M. Kast, Designing molecular complexes using free-energy derivatives from liquid-state integral equation theory. *J. Phys. Condens. Matter.* **28**, 344004 (2016).

73. P. G. Polishchuk, T. I. Madzhidov, A. Varnek, Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided. Mol. Des.* **27**, 675–679 (2013).

74. B. Munos, Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discov.* **8**, 959–968 (2009).

75. D. W. Borhani, D. E. Shaw, The future of molecular dynamics simulations in drug discovery. *J. Comput. Aided. Mol. Des.* **26**, 15–26 (2012).

76. G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe, Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–95 (2014).

77. C. F. Wong, Systematic sensitivity analyses in free energy perturbation calculations. *J. Am. Chem. Soc.* **113**, 3208–3209 (1991).

78. P. R. Gerber, A. E. Mark, W. F. van Gunsteren, An approximate but efficient method to calculate free energy trends by computer simulation. Application to dihydrofolate reductase inhibitor complexes. *J. Camd.* **7**, 305–323 (1993).

79. D. Pearlman, Free energy derivatives: A new method for probing the convergence problem in free energy calculations. *J. Comput. Chem.* **15**, 105–123 (1994).

80. P. Cieplak, D. a. Pearlman, P. a. Kollman, Walking on the free energy hypersurface of the 18-crown-6 ion system using free energy derivatives. *J. Chem. Phys.* **101**, 627 (1994).

81. P. Cieplak, P. Kollman, A technique to study molecular recognition in drug design: preliminary application of free energy derivatives to inhibition of a malarial cysteine protease. *J. Mol. Recognit.* **9**, 103–112 (1996).

82. S. Francisco, P. A. Kollman, Y.-P. Pang, Applications of free energy derivatives to analog design. *Perspect. Drug Discov. Des.* **3**, 106–122 (1995).

83. P. E. Smith, W. F. Van Gunsteren, Predictions of free energy differences from a single simulation of the initial state. *J. Chem. Phys.* **100**, 577 (1994).

84. H. Liu, A. E. Mark, W. F. van Gunsteren, Estimating the Relative Free Energy of Different Molecular States with Respect to a Single Reference State. *J. Phys. Chem.* **100**, 9485–9494 (1996).

85. L.-P. Lee, B. Tidor, Optimization of electrostatic binding free energy. *J. Chem. Phys.* **106**, 8681 (1997).

86. E. Kangas, B. Tidor, Charge optimization leads to favorable electrostatic binding free energy. *Phys. Rev. E.* **59**, 5958–5961 (1999).

87. P. a Sims, C. F. Wong, J. A. McCammon, Charge optimization of the interface between protein kinases and their ligands. *J. Comput. Chem.* **25**, 1416–29 (2004).

88. M. K. Gilson, Sensitivity Analysis and Charge-Optimization for Flexible Ligands: Applicability to Lead Optimization. *J. Chem. Theory Comput.* **2**, 259–270 (2006).

89. Y. Shen, M. K. Gilson, B. Tidor, Charge Optimization Theory for Induced-Fit Ligands. *J. Chem. Theory Comput.* **8**, 4580–4592 (2012).

90. H. C. Andersen, D. Chandler, Optimized Cluster Expansions for Classical Fluids. I. General Theory and Variational Formulation of the Mean Spherical Model and Hard Sphere Percus-Yevick Equations. *J. Chem. Phys.* **57**, 1918 (1972).

91. D. Chandler, H. C. Andersen, Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids. *J. Chem. Phys.* **57**, 1930 (1972).

92. D. Beglov, B. Roux, An integral equation to describe the solvation of polar molecules in liquid water. *J. Phys. Chem. B.* **101**, 7821–7826 (1997).

93. A. Kovalenko, F. Hirata, Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach. *Chem. Phys. Lett.* **290**, 237–244 (1998).

94. A. Kovalenko, F. Hirata, Self-consistent description of a metal--water interface by the Kohn--Sham density functional theory and the three-dimensional reference interaction site model. *J. Chem. Phys.* **110**, 10095–10112 (1999).

95. M. Ikeguchi, J. Doi, Direct numerical solution of the Ornstein–Zernike integral equation and spatial distribution of water around hydrophobic molecules. *J. Chem. Phys.* **103**, 5011 (1995).

96. D. Beglov, B. Roux, Numerical solution of the hypernetted chain equation for a solute of arbitrary geometry in three dimensions. *J. Chem. Phys.* **103**, 360 (1995).

97. A. G. Leach *et al.*, Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **49**, 6672–6682 (2006).

98. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016).

99. T. Chen, C. Guestrin, XGBoost. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16*, 785–794 (2016).

100. J. C. Gumbart, B. Roux, C. Chipot, Standard binding free energies from computer simulations: What is the best strategy? *J. Chem. Theory Comput.* **9**, 794–802 (2013).

101. Y. Deng, B. Roux, Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theory Comput.* **2**, 1255–1273 (2006).

102. J. Wang, Y. Deng, B. Roux, Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys. J.* **91**, 2798–814 (2006).

103. M. K. Gilson, J. A. Given, B. L. Bush, J. A. McCammon, The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **72**, 1047–1069 (1997).

104. C. Yung-Chi, W. H. Prusoff, Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* **22**, 3099–3108 (1973).

105. C. Bissantz, B. Kuhn, M. Stahl, A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **53**, 5061–5084 (2010).

106. A. S. Mahadevi, G. N. Sastry, Cooperativity in Noncovalent Interactions. *Chem. Rev.* **116**, 2775–2825 (2016).

107. E. Persch, O. Dumele, F. Diederich, Molecular Recognition in Chemical and Biological Systems. *Angew. Chemie - Int. Ed.*, 3290–3327 (2015).

108. A. Bauzá, T. J. Mooibroek, A. Frontera, The Bright Future of Unconventional σ/π-Hole Interactions. *ChemPhysChem*. **16**, 2496–2517 (2015).

109. P. Setny, Hydrophobic interactions between methane and a nanoscopic pocket: Three dimensional distribution of potential of mean force revealed by computer simulations. *J. Chem. Phys.* **128**, 125105 (2008).

110. P. Setny, Water properties and potential of mean force for hydrophobic interactions of methane and nanoscopic pockets studied by computer simulations. *J. Chem. Phys.* **127**, 54505 (2007).

111. G. Hummer, Molecular binding: Under water's influence. *Nat. Chem.* **2**, 906–907 (2010).

112. P. Setny, R. Baron, J. A. McCammon, How Can Hydrophobic Association Be Enthalpy Driven? *J Chem Theory Comput*. **6**, 2866–2871 (2010).

113. Á. Tarcsay, G. M. Keseru, Is there a link between selectivity and binding thermodynamics profiles? *Drug Discov. Today*. **20**, 86–94 (2015).

114. T. Young *et al.*, Dewetting transitions in protein cavities. *Proteins Struct. Funct. Bioinforma.* **78**, 1856–1869 (2010).

115. K. E. Rogers *et al.*, On the Role of Dewetting Transitions in Host–Guest Binding Free Energy Calculations. *J. Chem. Theory Comput.* **9**, 46–53 (2013).

116. J.-P. Hansen, I. R. I. . McDonald, *Theory of simple liquids* (1990).

117. T. Morita, K. Hiroike, A New Approach to the Theory of Classical Fluids. IIIa. *Prog. Theor. Phys.* **25**, 537–578 (1961).

118. L. S. Ornstein, F. Zernike, Integral equation in liquid state theory. *Proc. Acad. Sci. Amsterdam*. **17** (1914).

119. M. Llano-Restrepo, W. G. Chapman, Bridge function and cavity correlation function from simulation: Implications on closure relations. *Int. J. Thermophys.* **16**, 319–326 (1995).

120. I. Vyalov, G. Chuev, N. Georgi, Solute-solvent cavity and bridge functions. I. Varying size of the solute. *J. Chem. Phys.* **141**, 74505 (2014).

121. G. N. Chuev, I. Vyalov, N. Georgi, Extraction of atom–atom bridge and direct correlation functions from molecular simulations: A test for ambient water. *Chem. Phys. Lett.* **561–562**, 175–178 (2013).

122. L. Blum, A. J. Torrula, Invariant Expansion for Two-Body Correlations: Thermodynamic Functions, Scattering, and the Ornstein—Zernike Equation. *J. Chem. Phys.* **56**, 303–310 (1972).

123. J. Richardi, C. Millot, P. H. Fries, A molecular Ornstein–Zernike study of popular models for water and methanol. *J. Chem. Phys.* **110**, 1138–1147 (1999).

124. P. Jedlovszky, J. Richardi, Comparison of different water models from ambient to supercritical conditions: A Monte Carlo simulation and molecular Ornstein–Zernike study. *J. Chem. Phys.* **110**, 8019–8031 (1999).

125. K. Hiroike, On the theory of fluids. *J. Phys. Soc. Japan*. **13**, 1497–1503 (1958).

126. E. Meeron, Theory of Potentials of Average Force and Radial Distribution Functions in Ionic Solutions. *J. Chem. Phys.* **28**, 630–643 (1958).

127. S. M. Kast, T. Kloss, Closed-form expressions of the chemical potential for integral equation closures with certain bridge functions. *J. Chem. Phys.* **129**, 236101 (2008).

128. S. Bernèche, B. Roux, Molecular dynamics of the KcsA K(+) channel in a bilayer membrane. *Biophys. J.* **78**, 2900–17 (2000).

129. F. Hirata, The interionic potential of mean force in a molecular polar solvent from an extended RISM equation. *J. Chem. Phys.* **78**, 4133 (1983).

130. J. Heil, S. M. Kast, 3D RISM theory with fast reciprocal-space electrostatics. *J. Chem. Phys.* **142**, 114107 (2015).

131. E. O. Brigham, R. E. Morrow, The fast Fourier transform. *IEEE Spectr.* **4**, 63–70 (1967).

132. P. Duhamel, M. Vetterli, Fast fourier transforms: A tutorial review and a state of the art. *Signal Processing.* **19**, 259–299 (1990).

133. A. Rahman, Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.* **136**, A405–A411 (1964).

134. K. Lindorff-Larsen, P. Maragakis, S. Piana, D. E. Shaw, *J. Phys. Chem. B*, in press, doi:10.1021/acs.jpcb.6b02024.

135. K. G. Sprenger, V. W. Jaeger, J. Pfaendtner, The general AMBER force field (GAFF) can accurately predict thermodynamic and transport properties of many ionic liquids. *J. Phys. Chem. B.* **119**, 5882–5895 (2015).

136. J. W. Ponder *et al.*, Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B.* **114**, 2549–2564 (2010).

137. K. Chenoweth, A. C. T. van Duin, W. A. Goddard, ReaxFF Reactive Force Field for Molecular Dynamics Simulations of Hydrocarbon Oxidation. *J. Phys. Chem. A.* **112**, 1040–1053 (2008).

138. A. C. T. van Duin, S. Dasgupta, F. Lorant, W. A. Goddard, ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A.* **105**, 9396–9409 (2001).

139. V. Loup, L. Verlet, Computer experiments on classical fluids. *Phys. Rev.* **i**, 98–103 (1967).

140. W. C. Swope, A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **76**, 637 (1982).

141. R. W. Hockney, in *Methods in Computational Physics, Vol. 9* (1970), pp. 135–211.

142. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926 (1983).

143. H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).

144. J. L. Abascal, C. Vega, A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **123**, 234505 (2005).

145. C. Vega, J. L. F. Abascal, M. M. Conde, J. L. Aragones, What ice can teach us about water interactions: a critical comparison of the performance of different water models. *Faraday Discuss.* **141**, 251–276 (2009).

146. C. Vega, J. L. F. Abascal, Simulating water with rigid non-polarizable models: a general perspective. *Phys. Chem. Chem. Phys.* **13**, 19663 (2011).

147. S. M. Kast, personnel communication (2016).

148. J. G. Kirkwood, Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **3**, 300 (1935).

149. J. Hermans, S. Shankar, The Free Energy of Xenon Binding to Myoglobin from

Molecular Dynamics Simulation. *Isr. J. Chem.* **27**, 225–227 (1986).

150. N. Hansen, W. F. Van Gunsteren, Practical aspects of free-energy calculations: A review. *J. Chem. Theory Comput.* **10**, 2632–2647 (2014).

151. B. Kuhn *et al.*, Prospective Evaluation of Free Energy Calculations for the Prioritization of Cathepsin L Inhibitors. *J. Med. Chem.* **60**, 2485–2497 (2017).

152. Y. Miao, J. A. McCammon, Unconstrained enhanced sampling for free energy calculations of biomolecules: a review. *Mol. Simul.* **42**, 1046–1055 (2016).

153. M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids* (1989).

154. S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning* (Cambridge University Press, Cambridge, 2014).

155. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature.* **521**, 436–444 (2015).

156. J. Schmidhuber, Deep learning in neural networks: An overview. *Neural Networks.* **61**, 85–117 (2015).

157. J. Behler, Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **13**, 17930–17955 (2011).

158. J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 1–4 (2007).

159. J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 74106 (2011).

160. N. Artrith, T. Morawietz, J. Behler, High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B.* **83**, 153101 (2011).

161. R. M. Balabin, E. I. Lomakina, Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. *J. Chem. Phys.* **131**, 74104 (2009).

162. W. S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).

163. K. Murphy, *Machine Learning: a Probabilistic Perspective* (2012).

164. F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–408 (1958).

165. D. E. Rumelhart, P. Smolensky, J. L. McClelland, G. E. Hinton, in *Readings in Cognitive Science* (Elsevier, 1988), pp. 224–249.

166. C. M. Bishop, *Pattern Recognition and Machine Learning* (2006), vol. 4.

167. G. E. Hinton, S. Osindero, Y.-W. Teh, A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.* **18**, 1527–1554 (2006).

168. M. Abadi *et al.*, in *12th USENIX Symposium on Operating Systems* (2016; http://arxiv.org/abs/1605.08695).

169. J. Bergstra *et al.*, Theano: a CPU and GPU math compiler in Python. *Proc. Python Sci. Comput. Conf.*, 1–7 (2010).

170. F. Bastien *et al.*, Theano: new features and speed improvements. *arXiv:1211.5590*, 1–10 (2012).

171. The Theano Development Team *et al.*, Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints.* **abs/1605.0**, 19 (2016).

172. Y. Jia *et al.*, Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv Prepr. arXiv1408.5093* (2014).

173. X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks. *Proc. 13th Int. Conf. Artif. Intell. Stat.* **9**, 249–256 (2010).

174. V. Nair, G. E. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines. *Proc. 27th Int. Conf. Mach. Learn.*, 807–814 (2010).

175. K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition? *2009 IEEE 12th Int. Conf. Comput. Vis.*, 2146–2153 (2009).

176. D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. *Int. Conf. Learn. Represent. 2015*, 1–15 (2014).

177. H. M. Ashtawy, N. R. Mahapatra, A Comparative Assessment of Ranking Accuracies of Conventional and Machine-Learning-Based Scoring Functions for Protein-Ligand Binding Affinity Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **9**, 1301–1313 (2012).

178. J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29**, 1189–1232 (2001).

179. R. E. Schapire, in *Nonlinear Estimation and Classification*, D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, B. Yu, Eds. (Springer New York, New York, NY, 2003; http://dx.doi.org/10.1007/978-0-387-21579-2_9), pp. 149–171.

180. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. *Springer 2001*. **18**, 746 (2009).

181. J. Elith, J. R. Leathwick, T. Hastie, A working guide to boosted regression trees. *J. Anim. Ecol.* **77**, 802–813 (2008).

182. M. J. Frisch *et al.*, *Gaussian 03* (Gaussian, Inc., Wallingford, CT, 2003).

183. W. L. Jorgensen, J. Tirado-Rives, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657–1666 (1988).

184. W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118**, 11225–11236 (1996).

185. K. F. Schmidt, S. M. Kast, Hybrid Integral Equation/Monte Carlo Approach to Complexation Thermodynamics. *J. Phys. Chem. B.* **106**, 6289–6297 (2002).

186. A. D. MacKerell, N. Banavali, N. Foloppe, Development and current status of the CHARMM force field for nucleic acids. *Biopolymers.* **56**, 257–65 (2001).

187. I. S. Joung, T. E. Cheatham, Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B.* **112**, 9020–9041 (2008).

188. J. Aqvist, Ion-Water Interaction Potentials Derived from Free Energy Perturbation Simulations. *J. Phys. Chem.* **94**, 8021–8024 (1990).

189. G. Wipff, P. Weiner, P. Kollman, A molecular-mechanics study of 18-crown-6 and its alkali complexes: an analysis of structural flexibility, ligand specificity, and the macrocyclic effect. *J. Am. Chem. Soc.* **104**, 3249–3258 (1982).

190. J. Perkyns, B. M. Pettitt, A site–site theory for finite concentration saline solutions. *J. Chem. Phys.* **97**, 7656 (1992).

191. J. S. Perkyns, B. M. Pettitt, A dielectrically consistent interaction site theory for solvent—electrolyte mixtures. *Chem. Phys. Lett.* **190**, 626–630 (1992).

192. D. A. Case *et al.*, *AMBER12* (University of California, San Francisco, 2012).

193. J. C. Phillips *et al.*, Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).

194. H. J. C. Berendsen, D. van der Spoel, R. van Drunen, GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).

195. B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4**, 435–447 (2008).

196. S. Pronk *et al.*, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics.* **29**, 845–854 (2013).

197. Q. Du, D. Beglov, B. Roux, Solvation Free Energy of Polar and Nonpolar Molecules in Water: An Extended Interaction Site Integral Equation Theory in Three Dimensions. *J. Phys. Chem. B.* **104**, 796–805 (2000).

198. S. M. Kast, K. Friedemann Schmidt, B. Schilling, Integral equation theory for correcting truncation errors in molecular simulations. *Chem. Phys. Lett.* **367**, 398–404 (2003).

199. C. Hölzl *et al.*, Design principles for high–pressure force fields: Aqueous TMAO solutions from ambient to kilobar pressures. *J. Chem. Phys.* **144**, 144104 (2016).

200. *The PyMOL Molecular Graphics System* (Schrödinger, LLC, http://pymol.org, 2014).

201. T. Williams, C. Kelley, *Gnuplot 4.5: an interactive plotting program* (http://gnuplot.info, 2011).

202. *Mathematica 9.0* (Wolfram Research, Inc., 2012).

203. E. Glendening, An ab initio investigation of the structure and alkali metal cation selectivity of 18-crown-6. *J. Am.* …. **116**, 10657–10669 (1994).

204. D. Feller, Ab initio study of M+: 18-crown-6 microsolvation. *J. Phys. Chem. A.* **5639**, 2723–2731 (1997).

205. J. Rodriguez, T. Vaden, J. Lisy, Infrared spectroscopy of ionophore-model systems: hydrated alkali metal ion 18-crown-6 ether complexes. *J. Am. Chem. Soc.* **131**, 17277–17285 (2009).

206. J. Stark, Modellierung und Simulation von RET-Kinase-Komplexen, Technische Universität Dortmund (2015).

207. A. Šali, T. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

208. J. M. Wang, R. M. Wolf, J. W. Caldwell, P. a Kollman, D. a Case, Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

209. J. Wang, W. Wang, P. A. Kollman, D. A. Case, Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25**, 247–260 (2006).

210. A. Jakalian, D. B. Jack, C. I. Bayly, Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **23**, 1623–1641 (2002).

211. A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly, Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **21**, 132–146 (2000).

212. G. Sigalov, A. Fenley, A. Onufriev, Analytical electrostatics for biomolecules: Beyond the generalized Born approximation. *J. Chem. Phys.* **124**, 1–14 (2006).

213. D. A. Case *et al.*, *AMBER14* (University of California, San Francisco, 2014).

214. S. C. Hoops, K. W. Anderson, K. M. Merz, Force field design for metalloproteins. *J. Am. Chem. Soc.* **113**, 8262–8270 (1991).

215. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

216. U. Essmann *et al.*, A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577 (1995).

217. S. Miyamoto, P. A. Kollman, Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).

218. E. A. Meyer, M. Furler, F. Diederich, R. Brenk, G. Klebe, Synthesis and In Vitro Evaluation of 2-Aminoquinazolin-4(3H)-one-Based Inhibitors for tRNA-Guanine Transglycosylase (TGT). *Helv. Chim. Acta.* **87**, 1333–1356 (2004).

219. R. Frach, S. Kast, Solvation Effects on Chemical Shifts by Embedded Cluster Integral Equation Theory. *J. Phys. Chem. A* (2014).

220. Y. Alber, Fluor-Substitutionseffekt auf Protein-Ligand-Bindungsaffinitäten, Technische Universität Dortmund (2016).

221. *RDKit: Open-source cheminformatics* (http://www.rdkit.org, 2016; http://www.rdkit.org).

222. J. A. Maier *et al.*, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

223. P. Li, B. P. Roberts, D. K. Chakravorty, K. M. Merz, Rational design of particle mesh ewald compatible lennard-jones parameters for +2 metal cations in explicit solvent. *J. Chem. Theory Comput.* **9**, 2733–2748 (2013).

224. L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, A. Talwalkar, Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *arXiv Prepr.* (2016).

225. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfittin. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

226. F. Chollet, *Keras* (GitHub, https://github.com/fchollet/keras, 2015).

227. O. Russakovsky *et al.*, ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).

228. V. Schnecke, L. A. Kuhn, Virtual screening with solvation and ligand-induced complementarity. *Perspect. Drug Discov. Des.* **20**, 171–190 (2000).

229. DeepMind, AlphaGo (2017), (available at https://deepmind.com/research/alphago/).

230. V. Mnih *et al.*, Playing Atari with Deep Reinforcement Learning. *arXiv Prepr.*, 1–9 (2013).

231. V. Mnih *et al.*, Human-level control through deep reinforcement learning. *Nature.* **518**, 529–533 (2015).

232. H. van Hasselt, A. Guez, D. Silver, Deep Reinforcement Learning with Double Q-learning. *Artif. Intell.* **230**, 173–191 (2015).

233. Z. Wang *et al.*, Dueling Network Architectures for Deep Reinforcement Learning. *IEEE*

Commun. Mag. **54**, 48–57 (2015).

234. I. J. Goodfellow, On distinguishability criteria for estimating generative models. *5th Int. Conf. Learn. Represent.* (2015), doi:10.1109/CVPR.2005.287.

235. M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN. *arXiv* (2017).

236. I. Durugkar, I. Gemp, S. Mahadevan, Generative Multi-Adversarial Networks. *arXiv*, 1–14 (2016).

237. E. Denton, S. Chintala, A. Szlam, R. Fergus, Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. *arXiv*, 1–10 (2015).

8 Appendix

8.1 Pseudocode for the evaluation of the renormalized η-function

```
#calculation of the analytical 1D k-space monopol potentials
Do i=1, dimension second partner
        quu = qtotal1*qu2(i)
        monoUU1D = uqk_pot()
        monoUU1D = monoUU1D * expk *-kT1*renormFactor
end do
Do i=1, dimension solvent
        quv = qtotal1*qu2(i)
        monoUV1D = uqk_pot()
        monoUV1D = monopolUV * expk *-kT1*renormFactor
end do
#calculation of the monopol potentials in 3D k-space
monoUV3D = ewaldsummation(dim1=1, dim2=nv, qTot1, qvv)
monoUU3D = ewaldsummation(dim1=1, dim2=1, qTot1, qu2)
monoUU3D = kT1 * monoUU3D * renormFactor
#calculation of the last term (4)
Do j = 1, dimension solvent
        Do i = 1, dimension partner 2
                monoHK = monoHK + (-kT1 * monoUV1D)*rho*hk1d
        end do
end do
#calculation of the third and fourth term in 1D k-space
monoHK = monoHK + monoUU1D
```

```
do j = 1, dimension partner 2
        monoHR = revfft1D(monoHK)
        monoHR3D = interpolation(monoHR)
end do
#first part of the second term
ck3d = ck3d + (kT1 * monoUV3D)
hk0k1d = extrapolate(hk1D)
do i = 1, dimension partner 2
        do j = 1, dimension solvent
                hk3D = interpolation(hk01D)
                etak = etak + ck3d * (rho * hk3D)
        end do
end do
#ewald summation for partner 1 and 2 followed by calculation
#the last steps in the calculation of Term 12
ulkUU = ewaldsummation(dim1=dimension partner 1, dim2=dimension partner 2, qu1, qu2)
ulkUU = ulkUU * kT1 * renormFactor – monoUU3D
etak = etak – ulkUU
etar = revfft3D(etak)
#subtract Term 34 of Term12
etar = etar - monoHR3D
```

8.2 FEDs and PMFs for TGT/amq

Table 10: Atomwise FEDs and PMF values for the $TGT^{CH3,MD}$/$amq^{CH3,MD}$ system.

atom	$\frac{\partial w}{\partial \sigma}$ (p. *holo*)	$\frac{\partial w}{\partial \sigma}$ (*apo*)	$\frac{\partial w}{\partial \varepsilon}$ (p. *holo*)	$\frac{\partial w}{\partial \varepsilon}$ (*apo*)	$\frac{\partial w}{\partial q}$ (p. *holo*)	$\frac{\partial w}{\partial q}$ (*apo*)	w (p. *holo*)	w (*apo*)
C1	-1.25	-1.25	-0.89	-0.91	167.17	-1.21	-1.51	-1.00
C2	-0.69	-1.17	-0.62	-0.82	164.09	-1.54	-0.73	-0.73
C9	-0.07	-0.52	-0.23	-0.37	164.13	-1.12	-0.21	0.01
C3	-0.62	-1.22	-0.66	-0.92	164.81	-0.18	-0.83	-1.03
N3	-1.74	-1.76	-0.65	-0.64	16.84	-1.40	-9.41	-0.55
C8	-0.17	-1.31	-0.63	-1.04	163.84	0.88	-2.01	-0.81
N2	4.00	-0.32	-0.10	-0.48	178.21	-0.27	-15.71	-1.37
N1	0.06	0.49	-0.55	-0.39	178.76	0.78	-11.34	-0.90
C7	2.49	1.41	0.29	-0.10	167.53	6.15	-0.02	2.36
O1	5.01	-4.31	-0.17	-0.83	144.58	1.19	-54.94	-5.19
C4	-1.57	-1.07	-1.02	-0.80	170.61	-0.82	-3.24	0.29
C5	-1.50	-1.51	-0.94	-0.94	166.66	-2.90	-1.17	-0.28
C6	-1.24	-1.50	-0.89	-0.94	166.67	-1.50	-2.00	0.07
H1	-0.19	-0.34	-0.91	-1.17	56.18	0.51	-0.35	-0.39
H2	0.00	-0.14	-0.19	-0.46	52.45	-1.52	0.05	0.16
H3	-0.43	0.05	-1.17	-0.45	55.02	0.90	-0.44	-0.46
H4	-0.03	0.36	-0.29	0.34	51.89	-1.32	0.43	0.66
H5	-0.40	0.51	-0.50	0.61	101.49	-2.65	-1.76	-0.84
H6	-2.36	-1.77	-3.35	-2.72	99.60	-4.15	-2.91	-2.57
H7	-2.16	-1.64	-3.11	-2.51	88.14	-4.66	-1.83	-2.08
H8	-0.82	0.04	-2.04	-0.70	60.18	0.07	-0.78	-0.41
H9	-0.69	-0.47	-1.87	-1.61	62.06	-0.73	-0.39	-0.08

$\frac{\partial w}{\partial \sigma}$: kcal $mol^{-1} \cdot Å^{-1}$

$\frac{\partial w}{\partial q}$: kcal $mol^{-1} \cdot e^{-1}$

w: kcal mol^{-1}

Table 11: Atomwise FEDs and PMF values for the TGT^{H}/amq^{H} system.

atom	$\frac{\partial w}{\partial \sigma}$ (p. *holo*)	$\frac{\partial w}{\partial \sigma}$ (*apo*)	$\frac{\partial w}{\partial \varepsilon}$ (p. *holo*)	$\frac{\partial w}{\partial \varepsilon}$ (*apo*)	$\frac{\partial w}{\partial q}$ (p. *holo*)	$\frac{\partial w}{\partial q}$ (*apo*)	w (p. *holo*)	w (*apo*)
C1	-1.15	-2.38	-1.17	-1.48	53.61	-0.89	-2.94	-1.73
C2	-0.62	-1.39	-0.91	-1.10	50.67	-0.40	-1.51	-1.32
C3	-1.17	-1.51	-0.98	-1.00	48.15	-1.05	-1.67	-0.40
C4	-0.99	-1.54	-0.98	-1.15	50.33	0.38	-0.86	-1.10
N5	4.41	3.11	0.09	-0.02	35.94	1.90	-12.02	-1.70
C6	-0.60	-1.48	-1.04	-1.29	49.75	3.78	-0.93	0.48
N7	7.41	1.93	0.17	-0.32	67.99	1.71	-21.87	-3.66
N8	-1.13	-0.94	-0.78	-0.58	66.19	1.60	-14.51	-1.00
C9	1.93	1.22	0.24	-0.01	53.03	4.06	0.54	1.58
O10	2.77	-2.57	-0.14	-0.48	70.53	-0.58	-36.25	-2.32
C11	-1.64	-0.91	-1.14	-0.79	53.32	-0.68	-3.60	0.03
C12	-0.89	-1.93	-0.86	-1.16	51.66	-1.48	-1.29	-0.63
H1	3.49	3.41	5.63	5.57	16.54	2.29	1.01	0.60
H2	2.25	1.86	3.09	2.49	14.01	2.33	0.62	0.34
H3	-0.12	0.05	-0.73	-0.42	22.21	0.24	0.08	0.62
H4	-2.29	-2.57	-3.37	-3.72	87.52	-3.30	-2.38	-2.70
H5	-2.52	-2.27	-3.48	-3.27	87.90	-0.59	-2.64	-1.67
H6	-2.76	-1.99	-3.79	-2.94	63.08	-1.72	-0.74	-1.03
H7	0.24	0.44	-0.39	-0.11	14.95	5.23	0.66	0.92

$\frac{\partial w}{\partial \sigma}$: kcal $mol^{-1} \cdot Å^{-1}$

$\frac{\partial w}{\partial q}$: kcal $mol^{-1} \cdot e^{-1}$

w: kcal mol^{-1}

Table 12: Atomwise FEDs and PMF values for the TGT^{CH3}/amq^{CH3} system.

atom	$\frac{\partial w}{\partial \sigma}$ (p. *holo*)	$\frac{\partial w}{\partial \sigma}$ (*apo*)	$\frac{\partial w}{\partial \varepsilon}$ (p. *holo*)	$\frac{\partial w}{\partial \varepsilon}$ (*apo*)	$\frac{\partial w}{\partial q}$ (p. *holo*)	$\frac{\partial w}{\partial q}$ (*apo*)	w (p. *holo*)	w (*apo*)
C1	-1.08	-0.71	-0.96	-0.86	139.02	-0.07	-1.81	-1.06
C2	-0.94	-1.30	-0.82	-0.96	136.93	-1.58	-1.16	-0.74
C9	-0.33	-1.49	-0.47	-0.81	136.64	-1.02	-0.64	-0.13
C3	-0.75	-1.24	-0.82	-0.99	137.55	-0.22	-0.54	-1.06
N3	-1.18	-3.26	-0.85	-1.02	14.47	-1.23	-2.41	-1.66
C8	0.04	-1.19	-0.71	-1.12	136.68	2.84	-0.48	0.45
N2	0.78	-3.31	-0.83	-1.12	149.84	-0.70	-19.65	-2.65
N1	-0.29	-0.35	-0.68	-0.53	157.63	2.09	-16.83	-1.57
C7	2.89	1.33	0.52	-0.07	139.26	6.12	1.46	2.48
O1	2.53	-4.28	-0.30	-0.77	111.02	-0.45	-43.88	-3.89
C4	-1.86	-0.65	-1.12	-0.63	143.71	-0.85	-3.94	0.55
C5	-1.80	-1.94	-1.07	-1.12	139.78	-2.42	-1.34	-0.15
C6	-1.70	-2.26	-1.14	-1.31	136.59	-1.29	-2.15	-0.10
H1	2.19	2.13	3.08	2.92	42.85	2.32	0.73	0.44
H2	-0.26	0.25	-0.72	0.07	49.45	-2.23	0.14	0.95
H3	-0.09	0.14	-1.15	-0.91	49.17	1.06	-0.84	-0.97
H4	-1.21	-0.71	-2.72	-1.99	48.72	-0.55	-0.66	-0.01
H5	-2.48	-2.38	-3.50	-3.56	98.36	-3.70	-2.78	-2.37
H6	-2.50	-2.17	-3.44	-3.13	99.65	-0.68	-2.42	-1.35
H7	-2.53	-1.86	-3.56	-2.78	77.04	-2.83	-0.67	-1.64
H8	-1.11	-0.30	-2.78	-1.61	50.97	3.96	-0.05	0.01
H9	-1.19	-0.62	-2.91	-2.05	50.02	-1.97	-1.16	-0.44

$\frac{\partial w}{\partial \sigma}$: kcal $mol^{-1} \cdot Å^{-1}$

$\frac{\partial w}{\partial q}$: kcal $mol^{-1} \cdot e^{-1}$

w: kcal mol^{-1}

8.3 FED visualisation for the TGT ^{CH3}/amq^{CH3} system

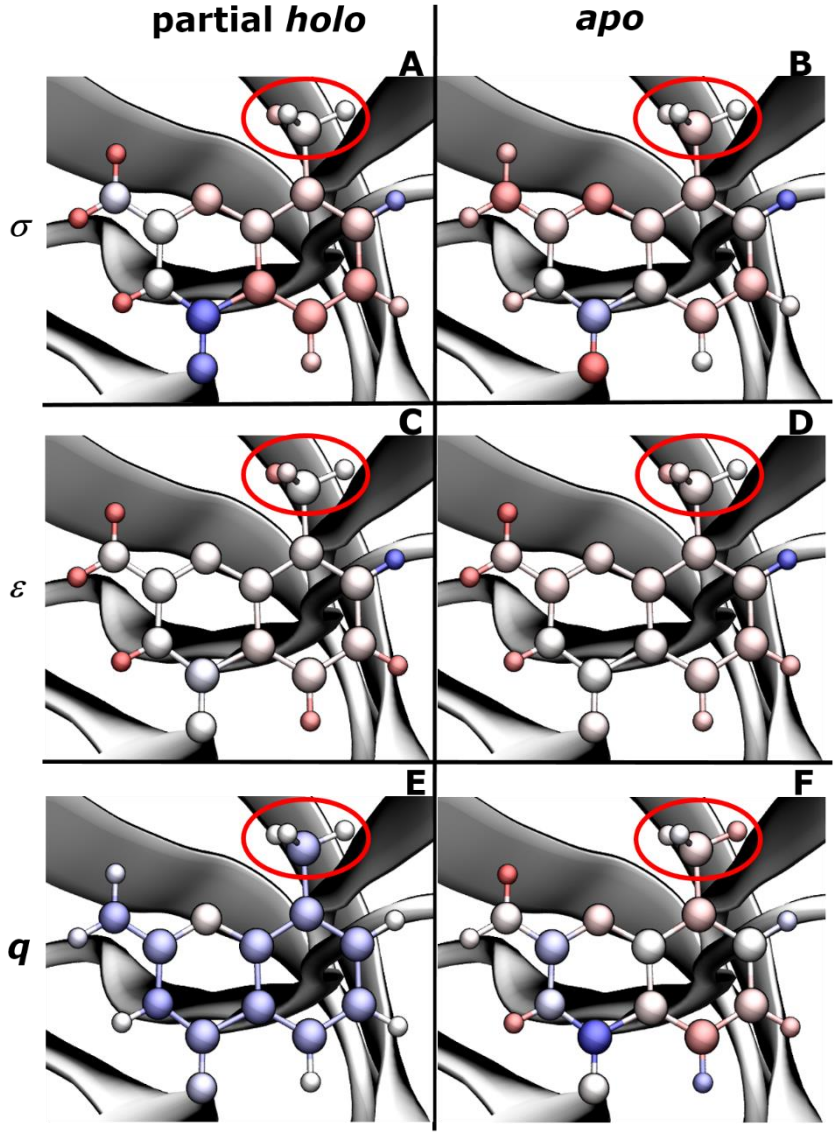

Figure 31: Atomwise FEDs for the TGT/amq^{CH3} system. The upper row shows the FEDs with respect to the σ value for the partial *holo* (A) and *apo* (B) calculations. In the middle row the FEDs with respect to the Lennard-Jones Parameter ε are shown for the partial *holo* (C) and *apo* (D) are shown. The last row shows the FEDs in regard to the partial charge q for the partial *holo* (E) and *apo* (F) calculations. The atoms are colour coded from red to white up to blue with red associated with a negative FED value and blue with a positive FED value. The atom group of particular interest is encircled in red.

8.4 Prediction data for the scoring functions

Table 13: Prediction data for the whole core dataset. F: fingerprint data only. F + P: fingerprint + PMF data. Predictions with the highest difference (negative red, positive green) to the experimental data are highlighted in bold.

		DNN1		DNN2		XGBoost	
PDB	**exp. data**	**F**	**F + P**	**F**	**F + P**	**F**	**F + P**
1r5y	6.46	5.60	6.18	4.95	5.87	5.40	4.80
1sln	6.64	5.32	5.38	5.40	4.92	6.06	5.72
2x8z	7.96	4.20	5.98	4.82	6.26	5.41	5.49
2yki	9.46	9.79	9.42	8.60	9.17	6.54	6.74
3k5v	6.30	5.68	5.95	5.69	6.41	5.91	6.18
3ivg	4.30	6.04	6.11	6.65	6.36	5.93	5.94
1o5b	5.77	5.62	5.98	6.12	5.93	5.45	5.18
3ao4	2.07	2.16	3.33	3.12	3.63	4.79	4.59
3pxf	4.43	5.03	5.11	5.36	5.29	4.99	5.25
1hfs	8.70	**3.37**	5.85	**4.59**	6.06	6.32	6.79
1mq6	11.15	7.53	**7.19**	7.87	**7.16**	**6.69**	**7.06**
1gpk	5.37	6.00	5.85	6.11	6.15	6.00	5.93
4djv	6.72	7.37	6.42	6.44	5.98	5.87	6.60
3uex	6.92	6.32	6.67	6.44	6.44	5.83	6.21
2y5h	5.79	6.55	7.60	7.77	7.03	6.78	7.08
3b3w	4.19	4.35	4.45	4.95	4.57	5.09	4.98
1bcu	3.28	6.28	5.66	5.93	5.52	5.33	4.78
1hnn	6.24	5.77	5.14	5.28	5.32	6.45	6.00
3zso	5.12	4.39	5.04	3.93	4.69	4.81	5.23
2qbp	8.40	8.47	6.70	8.14	6.93	6.50	6.49
3su2	7.35	6.32	6.92	7.31	7.00	6.62	6.65
3gcs	7.25	8.80	9.32	9.01	9.14	7.65	7.88
1oyt	7.24	5.41	6.95	6.85	6.52	6.96	6.54
3gy4	5.10	3.86	4.06	3.96	3.99	4.26	4.09
3b3s	2.55	4.35	4.75	4.95	4.72	5.09	5.07
3zsx	3.28	3.61	3.96	3.63	3.51	4.41	4.75
3imc	2.96	5.22	4.48	4.94	4.46	4.52	3.96
4de2	4.12	4.86	5.25	5.08	5.43	5.91	6.26
3f3c	6.02	3.89	4.34	4.47	4.40	5.63	5.69
2fvd	8.52	6.97	8.07	7.21	7.97	6.80	6.86
3ehy	5.85	5.57	5.29	5.39	5.48	5.38	4.94
2xnb	6.83	8.35	8.72	8.29	8.32	7.37	7.24
3mss	4.66	4.40	5.22	4.81	5.55	5.57	5.84
3huc	5.99	4.59	5.18	4.73	5.26	5.84	5.93
2gss	4.94	6.78	5.68	7.23	5.27	5.47	5.56
2p4y	9.00	6.27	7.75	6.32	7.91	6.90	7.17
3kgp	2.57	4.50	4.03	4.84	4.32	4.31	4.04
3su3	9.13	6.45	6.42	6.27	7.45	6.38	6.38
1p1q	4.89	7.12	6.37	6.64	6.40	6.50	6.13
1a30	4.30	5.71	5.89	6.74	5.38	5.37	5.72
3bfu	6.27	7.23	6.79	6.93	6.63	6.11	5.84
2qbr	6.33	8.18	7.29	7.81	7.51	5.86	6.23
4g8m	7.89	6.70	7.09	6.67	6.78	5.41	5.37
3gbb	6.90	4.56	5.46	5.02	5.53	5.64	5.42
3u9q	4.38	6.34	6.54	6.48	6.32	5.83	5.63
1uto	2.27	5.34	**4.73**	5.04	**4.89**	4.73	4.07
1sqa	9.21	6.68	7.37	7.30	7.19	6.65	6.69
3su5	5.58	6.45	6.58	6.27	7.58	6.38	6.35
3f17	8.63	6.52	5.41	6.36	5.49	5.63	5.33
3kwa	4.08	5.35	6.53	5.85	6.37	5.38	5.49
2hb1	3.80	6.57	5.73	6.60	5.71	4.88	4.88
3uo4	6.52	5.88	5.53	6.20	5.55	5.05	5.65
3g2z	2.36	**5.78**	4.47	**5.71**	4.68	**5.71**	**4.99**
2weg	6.50	6.94	7.12	6.78	7.05	6.24	6.42

Table 14: Prediction data for the K_d core dataset. F: fingerprint data only. F + P: fingerprint + PMF data. Predictions with the highest difference (negative red, positive green) to the experimental data are highlighted in bold.

		DNN1		DNN2		XGBoost	
PDB	**exp. data**	**F**	**F + P**	**F**	**F + P**	**F**	**F + P**
2yki	9.46	6.07	6.34	5.68	6.21	**5.55**	6.35
3k5v	6.30	6.01	6.87	5.72	6.57	5.38	5.84
3ivg	4.30	**8.20**	**8.03**	**7.73**	**7.64**	5.77	5.96
3ao4	2.07	3.02	3.73	3.43	3.87	4.39	**4.53**
3pxf	4.43	4.99	4.71	4.87	4.82	5.19	5.14
3uex	6.92	6.42	6.67	6.21	6.22	5.90	6.27
1bcu	3.28	6.46	6.39	6.38	6.05	5.35	4.32
3zso	5.12	4.91	4.75	4.45	4.71	4.22	4.51
3gcs	7.25	6.55	6.83	6.17	6.49	5.96	6.08
3gy4	5.10	4.07	4.26	4.09	4.22	4.57	4.35
3zsx	3.28	4.88	4.31	4.92	4.11	3.88	3.97
3imc	2.96	4.13	3.42	3.95	3.71	3.88	3.26
3mss	4.66	5.37	5.12	4.85	5.01	5.69	5.64
3huc	5.99	5.62	5.98	5.43	5.74	6.02	6.13
1p1q	4.89	7.78	7.54	7.42	7.32	6.56	6.19
4g8m	7.89	7.16	7.51	7.00	7.24	5.60	5.72
1uto	2.27	5.05	4.60	4.92	4.52	**4.61**	3.95
3f17	8.63	**4.61**	**4.07**	**4.71**	**4.23**	5.71	**5.22**
3uo4	6.52	7.39	6.54	7.24	6.86	6.08	6.03
2weg	6.50	6.56	6.82	6.58	6.92	6.07	6.21

Table 15: Prediction data for the K_i core dataset. F: fingerprint data only. F + P: fingerprint + PMF data. Predictions with the highest difference (negative red, positive green) to the experimental data are highlighted in bold.

		DNN1		DNN2		XGBoost	
PDB	**exp. data**	**F**	**F + P**	**F**	**F + P**	**F**	**F + P**
1r5y	6.46	5.89	6.09	5.89	6.04	5.55	4.90
1sln	6.64	4.96	3.69	5.11	3.98	5.69	5.79
2x8z	7.96	6.42	6.63	6.26	6.83	6.89	5.78
1o5b	5.77	5.80	6.13	5.77	6.19	5.70	5.48
1hfs	8.70	6.52	8.19	6.58	7.70	6.45	6.97
1mq6	11.15	8.10	7.47	8.41	7.48	**6.97**	**7.37**
1gpk	5.37	7.21	6.47	7.11	6.23	6.16	5.89
4djv	6.72	6.18	6.86	6.16	6.41	6.60	6.56
2y5h	5.79	6.96	6.98	6.71	7.21	6.85	7.16
3b3w	4.19	4.33	4.05	4.29	4.21	4.49	4.78
1hnn	6.24	6.43	5.90	6.41	5.59	6.14	6.27
2qbp	8.40	7.40	6.33	7.32	6.12	6.88	6.35
3su2	7.35	6.77	6.83	6.58	6.74	6.73	6.67
1oyt	7.24	6.67	6.00	6.23	6.31	6.89	6.32
3b3s	2.55	4.33	4.31	4.29	4.38	4.49	4.70
4de2	4.12	4.07	4.17	3.90	4.21	6.12	5.90
3f3c	6.02	5.07	4.57	5.03	4.71	5.23	5.20
2fvd	8.52	6.40	6.42	6.30	6.34	6.98	7.40
3ehy	5.85	5.34	4.81	5.39	4.78	5.57	5.43
2xnb	6.83	8.72	9.01	8.77	8.77	7.15	7.24
2gss	4.94	5.69	4.58	5.62	4.68	5.44	5.19
2p4y	9.00	10.93	12.07	11.03	10.80	6.76	7.35
3kgp	2.57	3.81	3.35	3.74	3.50	4.03	3.71
3su3	9.13	6.12	7.31	6.05	7.07	6.66	6.24
1a30	4.30	6.43	7.32	6.40	7.16	5.50	6.16
3bfu	6.27	6.18	6.10	6.19	5.97	4.89	5.30
2qbr	6.33	7.69	7.01	7.80	6.80	6.32	6.30
3gbb	6.90	5.37	6.53	5.38	6.52	5.84	5.54
3u9q	4.38	5.89	6.40	6.03	6.49	4.78	5.31
1sqa	9.21	**5.13**	**5.45**	**5.09**	**5.48**	6.21	6.80
3su5	5.58	6.12	7.11	6.05	6.84	6.66	6.34
3kwa	4.08	3.91	4.69	3.75	4.93	4.65	5.28
2hb1	3.80	6.01	5.45	5.92	5.24	5.09	4.78
3g2z	2.36	**6.62**	**5.81**	**6.80**	**6.13**	**5.60**	**5.10**

9 Electronic appendix