Local Thermodynamic Characterisation of Binding Sites and Protein-Ligand Interactions

Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) vorgelegt von Julia B. Jasper

Diese Dissertation wurde im Zeitraum vom 01.04.2017 bis zum 07.02.2022 an der Fakultät Chemie und Chemische Biologie der Technischen Universität Dortmund angefertigt

Erstgutachter: Prof. Dr. Stefan M. Kast
Zweitgutachter: Dr. Andreas Brunschweiger

Acknowledgement

Zuallererst möchte ich mich bei Prof. Dr. Stefan M. Kast bedanken. Zum einen natürlich für die Möglichkeit, an einem so spannenden Thema zu arbeiten, für den wissenschaftlichen Input und für die fachliche Unterstützung. Zum anderen aber auch ganz besonders für die angenehme Arbeitsatmosphäre und die Menschlichkeit und große Integrität auch in schwierigen Situationen. Ich werde mich immer gerne daran erinnern, wie gut ich in der Arbeitsgruppe aufgenommen wurde und wie wohl ich mich dort gefühlt habe.

Darüber hinaus möchte ich mich bei Dr. Andreas Brunschweiger bedanken – für die Übernahme der Zweitkorrektur und für eine sehr angenehme und inspirierende Zusammenarbeit während meiner gesamten Zeit in der Kast-Arbeitsgruppe. Die Kooperation mit der Arbeitsgruppe Brunschweiger hat meine Arbeit um viele spannende Beispiele und lehrreiche Diskussionen bereichert.

Bei der gesamten Arbeitsgruppe Kast möchte ich mich für die tolle Atmosphäre und viele wunderbare und lustige Momente bedanken. Ganz besonders honorieren möchte ich hierbei:

Patrick für Wortspiele und Witze nur auf dem allerhöchsten Niveau.

Martin für seine Sorge um unser kulinarisches Wohlergehen.

Meinen Büro-Nachbarn Tim und Yannic für sinnvolle und sinnfreie, jedoch fast immer unterhaltsame Diskussionen aller Art.

Lars für seine Geduld und seinen Humor bei der gemeinsamen Durchführung von Praktika und der Protokoll-Korrektur.

Außerdem möchte ich mich bei allen Freunden und Verwandten bedanken, die mich während der ganzen Zeit unterstützt haben. Besonderer Dank geht hier an Marina, Dani und Lydia für den tatkräftigen und emotionalen Support, an Nico und Chrizzy für zahlreiche hochintellektuelle Freitag- und Samstagabende, an Tim und Giesi für wunderbare Unternehmungen und Kurzurlaube, an Teresa und Linda für die tolle Gesellschaft während der Corona-Zeit sowie an Tobi und Lea, Michi und Lasse, Oli, Ina, Julia, Marius, Martin und Sonja und alle, die ich hier ggf. noch vergessen habe.

Zuletzt möchte ich mich bei Lukas bedanken, einfach für alles Bisherige und alles, was noch kommt.

Eidesstattliche Versicherung (Affidavit)

Jasper, Julia

Name, Vorname
(Surname, first name)

141576

Matrikel-Nr.
(Enrolment number)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden, § 63 Abs. 5 Hochschulgesetz NRW.

Die Abgabe einer falschen Versicherung an Eides statt ist strafbar.

Wer vorsätzlich eine falsche Versicherung an Eides statt abgibt, kann mit einer Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft werden, § 156 StGB. Die fahrlässige Abgabe einer falschen Versicherung an Eides statt kann mit einer Freiheitsstrafe bis zu einem Jahr oder Geldstrafe bestraft werden, § 161 StGB.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:

Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offences of this type is the chancellor of the TU Dortmund University. In the case of multiple or other serious attempts at deception, the candidate can also be unenrolled, Section 63, paragraph 5 of the Universities Act of North Rhine-Westphalia.

The submission of a false affidavit is punishable.

Any person who intentionally submits a false affidavit can be punished with a prison sentence of up to three years or a fine, Section 156 of the Criminal Code. The negligent submission of a false affidavit can be punished with a prison sentence of up to one year or a fine, Section 161 of the Criminal Code.

I have taken note of the above official notification.

Ort, Datum
(Place, date)

Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

Local Thermodynamic Characterisation of Binding Sites and Protein-Ligand Interactions

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.
Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder der TU Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

I hereby swear that I have completed the present dissertation independently and without inadmissible external support. I have not used any sources or tools other than those indicated and have identified literal and analogous quotations.

The thesis in its current version or another version has not been presented to the TU Dortmund University or another university in connection with a state or academic examination.*

*Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the PhD thesis is the official and legally binding version.

Ort, Datum
(Place, date)

Unterschrift
(Signature)

Table of Contents

1. Motivation and objectives

Abstract

With the advent of high-performance computing and progress in fields like structural biology, structure-based drug design (SBDD) has become an essential part in modern drug discovery projects. The knowledge about the basic physicochemical principles of molecular interactions, combined with three-dimensional structures of protein binding sites and small molecules, allows to make predictions about ligand binding modes and affinities. Over the past decades, various methods, ranging from simple descriptor-based approaches to the application of quantum mechanics and machine learning, have been developed that aim at identifying the optimal small molecule partner for a given protein target to achieve a desired biological modulation, and many of them have already been successfully applied in the development of lead structures and approved drugs.

Yet, the "holy grail" of rational drug discovery, the completely automated *de novo* design of the "perfect" ligand to a protein structure, so far remains unreached, even with today's (re)emergence of powerful machine learning techniques. One of the reasons for this is the difficulty to consider all thermodynamically relevant details of protein-ligand complex formation in the modelling and design process. In classical SBDD approaches, the intermolecular interactions between host and guest, like hydrogen bonds, van der Waals interactions, or salt bridges, are usually well accounted for. More subtle aspects include the change in both ligand and protein conformations upon binding and, above all, solvent effects.

Investigating these solvent effects, and especially deriving rules how to exploit knowledge about the local thermodynamic properties of protein hydration sites for drug design purposes, is the main objective of this work. To tackle this challenge, the method of choice used in this work is the 3D reference interaction site model (3D RISM) which is based on integral equation theory. It does not only allow to predict the solvent distribution within a binding site, but also to calculate local thermodynamic properties of specific hydration sites. The first part of this work therefore focuses on a large-scale analysis of the thermodynamic signatures of protein hydration sites and their correlation with ligand features. Water replacement rules for use in ligand design and optimisation are derived and exemplified on the basis of matched molecular pairs (MMPs).

The second part of this work expands the concept of 3D RISM-derived thermodynamic binding site characterisation to virtual probe sites mimicking specific functional groups whose distribution within a

binding site can be calculated by a 3D RISM solute-solute approach implemented in the Kast working group by F. Mrugalla. An advanced framework combining RISM-based binding site characterisation with automatised library preparation, docking, and scoring is established that allows to “convert” the probe densities to a selection of promising fragments or small molecules, thus making one step towards the “holy grail” of automated *de novo* ligand design.

In a third part, the afore-mentioned concepts are applied together onto three case studies from the challenging field of protein-protein interactions (PPIs) to illustrate the practicability of the developed approaches in real-life medicinal chemistry examples.

Zusammenfassung

Mit dem Aufkommen von Hochleistungsrechnern und Fortschritten in Bereichen wie der Strukturbiologie ist das strukturbasierte Wirkstoffdesign zu einem wesentlichen Bestandteil moderner Arzneimittelforschungsprojekte geworden. Das Wissen über die grundlegenden physikalisch-chemischen Prinzipien molekularer Wechselwirkungen, kombiniert mit dreidimensionalen Strukturen von Proteinbindetaschen und kleinen organischen Molekülen, ermöglicht Vorhersagen über die Bindemodi und Affinitäten von Liganden. In den letzten Jahrzehnten wurden verschiedene Methoden - von einfachen Deskriptor-basierten Ansätzen bis hin zur Anwendung von Quantenmechanik und maschinellem Lernen - entwickelt, die darauf abzielen, den optimalen Liganden für ein bestimmtes Proteintarget zu identifizieren, um eine gewünschte biologische Modulation zu erreichen. Viele von ihnen wurden bereits erfolgreich bei der Entwicklung von Leitstrukturen und zugelassenen Arzneimitteln eingesetzt.

Doch der "heilige Gral" der rationalen Arzneimittelforschung, das vollständig automatisierte *de novo* Design des "perfekten" Liganden für eine Proteinstruktur, bleibt bisher unerreicht, selbst mit den heute (wieder) aufkommenden Methoden des maschinellen Lernens. Einer der Gründe dafür ist die Schwierigkeit, alle thermodynamisch relevanten Details der Protein-Ligand-Komplexbildung im Modellierungs- und Designprozess zu berücksichtigen. In klassischen Ansätzen des strukturbasierten Wirkstoffdesigns werden die intermolekularen Wechselwirkungen zwischen Rezeptor und Ligand, wie Wasserstoffbrücken, van der Waals-Wechselwirkungen oder Salzbrücken, in der Regel gut berücksichtigt. Zu den subtileren Aspekten gehören die Änderung der Liganden- und Proteinkonformationen bei der Bindung und vor allem Lösungsmitteleffekte.

Die Untersuchung dieser Lösungsmitteleffekte und insbesondere die Ableitung von Regeln, wie das Wissen über die lokalen thermodynamischen Eigenschaften von Wassermolekülen in Proteinbindetaschen für die Entwicklung von Liganden genutzt werden kann, ist ein Hauptziel dieser Arbeit. Um diese Herausforderung zu bewältigen, wird in dieser Arbeit das 3D *reference interaction site model* (3D RISM) verwendet, das auf der Integralgleichungstheorie basiert. Es ermöglicht nicht nur die Vorhersage der Lösungsmittelverteilung innerhalb einer Bindetasche, sondern auch die Berechnung lokaler thermodynamischer Eigenschaften spezifischer Wassermoleküle. Der erste Teil dieser Arbeit konzentriert sich daher auf eine groß angelegte Analyse der thermodynamischen Signaturen von Wassermolekülen in Proteinbindetaschen und deren Korrelation mit Liganden-Eigenschaften. Es werden Regeln für die Verdrängung von Bindetaschen-Wassermolekülen durch Liganden abgeleitet, die

bei der Entwicklung und Optimierung von Liganden verwendet werden können; diese werden anhand von sog. *matched molecular pairs* (MMPs) veranschaulicht.

Im zweiten Teil dieser Arbeit wird das Konzept der thermodynamischen Charakterisierung von Bindetaschen auf virtuelle *probes* ausgeweitet, die spezifische funktionelle Gruppen nachahmen. Ihre Verteilung innerhalb einer Bindetasche kann durch einen 3D-RISM *uu*-Ansatz berechnet werden, der in der Kast-Arbeitsgruppe von F. Mrugalla implementiert wurde. Es wurde ein Workflow etabliert, der die RISM-basierte Charakterisierung von Bindetaschen mit automatisierter Bibliotheksvorbereitung, Docking und Scoring kombiniert und es ermöglicht, die *probe*-Dichten in eine Auswahl vielversprechender Fragmente oder kleiner Moleküle "umzuwandeln" - und damit einen Schritt in Richtung des "heiligen Grals" des automatisierten *de novo* Liganden-Designs macht.

In einem dritten Teil werden die oben genannten Konzepte auf drei Fallstudien aus dem anspruchsvollen Bereich der Protein-Protein-Interaktionen (PPIs) angewandt, um die Anwendbarkeit der entwickelten Ansätze in realen Beispielen der medizinischen Chemie zu veranschaulichen.

2. Introduction

2.1 Ligand-receptor binding

The basis of all biological processes, ranging from enzyme reactions to the transcription of DNA or hormone signalling, is the non-covalent formation of host-guest complexes between molecules. Understanding the thermodynamic and kinetic principles of molecular recognition is a prerequisite for the development of tools for modulating these interactions and, hence, physiological processes. Generally, the association of a complex LR from an arbitrary receptor R and a ligand L can be described via:[1]

$$\mathrm{L} + \mathrm{R} \underset{k_\mathrm{off}}{\overset{k_\mathrm{on}}{\rightleftharpoons}} \mathrm{LR}\,. \tag{1}$$

Here, k_on and k_off denote kinetic rate constants for the association and dissociation of the complex. In an equilibrium state, forward and backward reaction should be balanced, so that

$$k_\mathrm{on} c_\mathrm{R} c_\mathrm{L} = k_\mathrm{off} c_\mathrm{LR}\,, \tag{2}$$

with c_R, c_L, and c_LR as the equilibrium concentrations of free ligand, free receptor, and the respective complex (under the approximation that the respective activity constants are 1). From the ratio of k_on and k_off, the binding or association constant K_b can be defined according to:[1]

$$K_\mathrm{b} \equiv \frac{k_\mathrm{on}}{k_\mathrm{off}} \equiv \frac{1}{K_\mathrm{d}} \tag{3}$$

Its reciprocal is the dissociation constant K_d. K_d is given in units of concentration (i.e. mol/L) and is not a thermodynamic equilibrium constant. Hence, “good” binders show low K_d values (i.e. in the “nanomolar range”). In the context of protein ligand interactions, especially the inhibition of enzymes, often the inhibition constant K_i is reported, which simply denotes the special case of an equilibrium constant for the dissociation process of an inhibitor-enzyme complex.[2]

The binding constant of complex formation is related to the standard binding free energy, $\Delta_\mathrm{bind} G^\circ_\mathrm{LR}$, via the Gibbs relation, and can be expressed as the difference of the standard chemical potential $\mu^\circ_{\mathrm{sol},i}$ of the different species in solution via:[4]

$$\Delta_\mathrm{bind} G^\circ_\mathrm{LR} = -RT \ln\left(\frac{\gamma_\mathrm{LR}}{\gamma_\mathrm{L}\gamma_\mathrm{L}} \frac{C^0 C_\mathrm{LR}}{C_\mathrm{L} C_\mathrm{R}} \right) \equiv -RT \ln K_\mathrm{b}\,, \tag{4}$$

with the molar gas constant R, the absolute temperature T, γ_i as the activity constant of a given species in solution, and C^0 as the standard concentration. The more negative the associated change in the Gibbs free energy of the system is upon complex formation, the more favourable the reaction.[3] Under physiological conditions, not only ligand and receptor themselves are involved in this process, but also the surrounding solvent including buffer components and other molecules that are present in the solution (especially in living cells). This results in molecular crowding effects and a highly complex interplay of interactions and energy exchanges between all involved species.

Accordingly, $\Delta_{\text{bind}}G^{\circ}_{\text{LR}}$ can be calculated as the difference of the standard chemical potential $\mu^{\circ}_{\text{sol},i}$ of the different species in solution via:[4]

$$\Delta_{\text{bind}}G^{\circ}_{\text{LR}} \equiv \mu^{\circ}_{\text{sol,LR}} - \mu^{\circ}_{\text{sol,L}} - \mu^{\circ}_{\text{sol,R}} = -RT\ln\left(\frac{\gamma_{\text{LR}}}{\gamma_{\text{L}}\gamma_{\text{R}}}\frac{C^{\circ}C_{\text{LR}}}{C_{\text{L}}C_{\text{R}}}\right)_{eq} \equiv -RT\ln K_{\text{b}} \tag{5}$$

$\Delta_{\text{bind}}G^{\circ}_{\text{LR}}$ can be attributed to enthalpic and entropic contributions according to the fundamental relation:

$$\Delta_{\text{bind}}G^{\circ}_{\text{LR}} = \Delta_{\text{bind}}H^{\circ}_{\text{LR}} - T\Delta_{\text{bind}}S^{\circ}_{\text{LR}}\ , \tag{6}$$

with $\Delta_{\text{bind}}H^{\circ}_{\text{LR}}$ and $\Delta_{\text{bind}}S^{\circ}_{\text{LR}}$ denoting the (standard) enthalpic and entropic changes upon ligand binding. For simplicity, the standard notation is often (as in the later parts of this work) omitted but assumed implicitly. At the molecular level, these contributions can be attributed to several physical phenomena:[3] The most obvious is the formation of non-covalent interactions between the ligand and the receptor, including e.g. van der Waals interactions, hydrogen bonds, and salt bridges, which can be considered to be mainly enthalpy-driven. Effects dominated by entropy include changes in the configurational disorder of both binding partners upon association. Another important factor, which is a main topic of this work and is discussed in detail in the following sections, are changes in the solvation of ligand and receptor including the release of solvent molecules into bulk solvent or the formation or disruption of solvent-solvent or solvent-solute interactions. In this complex interplay, the changes in enthalpy and entropy are often opposed to each other, e.g. the formation of strong interactions between a ligand and a receptor is favourable from an enthalpic point but unfavourable in terms of entropy since it reduces the degrees of freedom for both compounds. This compensation effect is referred to as “enthalpy-entropy compensation”.[5]

2.2 Methods in SBDD

The general concept behind SBDD is to guide the design of a novel ligand by exploiting 3D structural information about its binding partner, e.g. a protein.[6] The very first attempts to explain molecular recognition on a structural basis date back to the 19th century, when E. Fischer proposed the “lock and key” concept according to which a complex can only be formed if the ligand (the key) perfectly matches the binding cavity (key hole) of the receptor.[7] In contrast to conventional forward drug discovery, where the identification of lead structures is obtained by experimental high-throughput screening (HTS), a rational design process is less costly and more effective since a ligand is especially tailored towards a known biological target.[8]

The foundation of SBDD lies in the progress in structural biology and computer science which enabled researchers to obtain, process, store and analyse the 3D structures of pharmaceutically relevant targets. Since the very first steps in the middle of the last century, when the X-ray structures of myoglobin and insulin were solved,[9,10] the field has massively advanced, resulting not only in several success-stories of FDA-approved drugs (e.g. Saquinavir,[11] Raltitrexed,[12] Amprenavir,[13] Isoniazid,[14], Epalrestat,[15] Flurbiprofen[16]) but also in the development of powerful algorithms and software, thus giving rise to methods like docking and molecular dynamics (MD) simulations. The detailed description of all of these applications is clearly without the scope of this work; yet, the basic ideas and principles behind the most important approaches will be presented in the following paragraphs in order to put the results of this work into context in the large menagerie of SBDD methods.

2.2.1 Binding site identification

Generally, the absolute prerequisite for the structure-based design of a ligand to a given protein is knowledge of the respective binding site. In case that only an *apo* structure, i.e. without a bound ligand, is available, it must be identified. A plethora of methods has been developed for this task which can be roughly grouped into sequence-based, template-based, geometric, and energy-based.[17] While the first two classes rely on the annotation of already known binding sites, the latter two in principle allow the identification of binding sites only from the structure without any similar templates.

The basic assumption of geometric approaches is that binding sites show a distinct shape, like a deep cleft or pocket, that can be distinguished from the rest of the protein surface by geometric descriptors. Prominent examples of algorithms include POCKET,[18] LIGSITE,[19] and SURFNET.[20]

The idea behind energy-based approaches is that binding sites exhibit specific energetic properties which can be identified. One of the very first applications used in this context is the famous GRID[21] approach by Goodford which calculates the interaction energy between the protein and chemical probes mimicking specific ligand functional groups, resulting in interaction maps. The GRID force fields is for instance used in the binding site identification program Q-SiteFinder,[22] but also other force fields have been employed by a variety of approaches.[23,24,25]

2.2.2 Lead identification by *de novo* design and virtual screening approaches

With the binding site identified, the next step in the SBDD process is finding a lead candidate. This can in principle be achieved by two different approaches, the complete *de novo* construction of a molecule, or hit identification in a virtual screening.

2.2.2.1 *De novo* design

For the *de novo* construction, building blocks in form of fragments, functional group units, or atoms, are fitted into the binding site. This is in analogy to the experimental fragment-based screening via co-crystallisation of fragments, which can then be grown or linked together via computational search algorithms and scoring functions. So far, however, few available ready-to-use software exists for this purpose. One of the first approaches developed for *de novo* design is LUDI[26] which calculates interaction sites based on the provided receptor using distributions of non-bonded contacts in the Cambridge Structural Database (CSD) and performs consequent fitting of fragments. The program SPROUT[27] utilises constraints in form of interaction sites derived from the steric, hydrophobic, and electrostatic properties of the desired binding partner to step by step generate a novel compound from user-defined templates. It has been applied to the design of inhibitors of β-Site APP cleaving enzyme 1 (BACE1).[28] The Program LigBuilder[29] constructs a ligand stepwise from a library of organic fragments under consideration of the given protein binding site by deriving a pharmacophore model based on the binding energies of the protein with different probes, similar to the GRID approach. It has been extended and used for the design of dual-target inhibitors of cyclooxygenase-2 (COX-2) and 5-lipoxygenase (5-LOX).[30] Besides, there are also ligand-based approaches that do not consider the target structure but rely on the structure of an already known ligand. The program DOGS (Design of Genuine Structures)[31] generates new compounds based on a single reference compound with desired activity and has been used in multiple case studies.[32,33,34] Besides, in recent years, machine-learning (ML)-based methods have been pursued, like the "generative artificial intelligence" approach coined by Schneider and co-workers that includes training of neural networks on databases of drug-like molecules.[35]

2.2.2.2 Virtual screening approaches

Much more well-established approaches exist for the identification of promising ligands via virtual screening, i.e. the selection of a compound with presumably desired biological properties from an existing *in silico* molecule library. As for the *de novo* design, two general approaches can be distinguished: ligand-based approaches, which utilise information about already known binders to i.e.

build a pharmacophore model or to carry out similarity searches, and receptor-based methods, which rely on the 3D structure of the binding partner and which will be discussed here in more detail.

2.2.2.2.1 Pharmacophores

One example for a receptor-based virtual screening method is pharmacophore-based screening. Although traditionally used in ligand-based virtual screening, pharmacophore models can also be derived from 3D ligand-receptor complex structures or even from *apo* receptor structures. The concept of a “pharmacophore” was already established in the beginning of the 20th century by P. Ehrlich as “a molecular framework that carries (“phoros”) the essential features responsible for a drug's (“pharmakon”) biological activity”.[36] Since then, the definition was extended towards “an arrangement of molecular features or structural elements related to biological activity”.[37,38] Commonly used features e.g. include hydrogen bond donor or acceptor (HBA, HBD), cation, anion, aromatic, or hydrophobic,[39,40] which can be defined based on suitable complementarity to structural features of an *apo* binding site or, if available, based on interactions in a known ligand-receptor complex structure. The advantage of deriving pharmacophore features from an *apo* binding site is that it can be applied for targets for which no ligands are known so far. Besides, even if there are known ligands, it can be beneficial to create an unbiased pharmacophore model completely independent from existing ligands to increase the chance of “scaffold hopping”,[41] i.e. finding compounds with so far unexploited scaffolds for improved synthetic availability or for circumventing problems with intellectual property.

Pharmacophore features can be derived in different ways, which can be generally differentiated into pattern-based and molecular field-based methods.[42] Pattern-based approaches e.g. assign predefined features like HBD based on the presence of functional groups. Molecular field-based methods utilise interaction fields which can be generated by GRID or similar applications to identify areas with highly favourable interaction energies between the binding site and specific molecular probes as location of pharmacophore features. A prominent example of such a field-based approach is FLAP.[43] Several programs and software packages have been developed for pharmacophore generation and pharmacophore-based screening that offer implementations of different feature assignment algorithms, i.e. MOE,[44] LigandScout,[45] PHASE,[46] and Catalyst.[47]

After generation of a pharmacophore model, virtual screening of *in silico* databases can be performed either based on fingerprint-comparison or by means of a 3D alignment. For fingerprint comparison, like used in FLAP, the information about the presence of pharmacophore features is converted to a fingerprint vector, thus allowing for efficient comparison by e.g. a Tanimoto coefficient.[48] In a 3D alignment, which is e.g. used by LigandScout, MOE and Catalyst, 3D conformations of the screening compounds are aligned to the 3D pharmacophore model to calculate the feature matching, which is computationally expensive.[42] A noteworthy peculiarity in this context is the pattern-matching 3D alignment algorithm of LigandScout which utilises inter-feature distance fingerprints to achieve a faster comparison when more features are included in the pharmacophore model, while usually the screening time drastically increases with the number of features.[49]

2.2.2.2.2 Docking

The probably most-frequently used approach for virtual screening is docking. Usually, the term "docking" is used to describe a two-step process, namely the prediction of the binding mode of a given compound in its receptor, typically a protein binding site, and the assessment of this pose via a scoring function to yield an approximation or relative measure of the respective binding affinity.[50]

Several well-established programs have been developed for conducting docking experiments, e.g. GOLD,[51] Glide,[52] Flexx,[53] DOCK,[54] and AutoDock,[55] each encompassing different algorithms for pose generation and scoring. In the following, a short overview will be given about the most common strategies concerning both the "docking" and "scoring" challenge.

Even under the (clearly over-simplified) assumption of a rigid ligand and receptor, which was used in the early program DOCK, docking can be considered a highly complex six-dimensional puzzle since a molecule can be translated and rotated within the binding site. When considering that both the ligand and the protein can undergo conformational changes upon binding, the challenge becomes even more complex, requiring elaborate and efficient algorithms to generate reasonable binding modes. Generally, there are two strategies, stochastic and systematic search, with the latter being differentiated into exhaustive search, fragmentation search and conformational ensemble search.[50] In the exhaustive search, as for instance used in Glide, all rotatable bonds are rotated systematically in certain increments to generate all possible conformations of the ligand, thus resulting in a combinatoric explosion for molecules with a large number of rotatable bonds. To reduce the computational cost, usually constraints based on the binding site are applied in the beginning of the pose generation process, and the exhaustive search is only carried out for the most promising conformations. In fragmentation search, as for instance used in Flexx, the ligands are divided into rigid fragments, and the binding conformation is grown inside the binding site by first docking a starting fragment and then adding the other ones. Conformational ensemble methods, on the other hand, perform rigid docking of a set of pre-computed ligand conformations.

In contrast to systematic search, stochastic search approaches aim at sampling the ligand conformational space inside the binding site by introducing random changes to both the angles of rotatable bonds and the positioning of the ligand. A change can be accepted or rejected according to a probabilistic criterion, for instance a Boltzmann probability function based on an estimate of the respective energy of the pose before and after the change. One of the most widely used docking programs, GOLD, utilises a genetic algorithm (GA). In a first step, a population of initial poses is generated via assignment of ligand dihedral angles, torsions, and ring conformations as well as positioning of the ligand inside the binding site via matching of HBD, HBA, and hydrophobic fitting points. Both the conformational parameters and the mapping of protein and ligand fitting points of a pose are stored in "chromosomes", and similar poses within the population are stored in "islands". In analogy to the biological evolution process, these initial "parent" poses are subjected to random mutation (change of individual values in a chromosome), cross-over (exchange of whole chromosome parts within islands), or migration (exchange of whole chromosomes between different islands). The probability of a specific chromosome being selected as parent chromosome, which can "inherit" its information to the next generation, is proportional to its assessment by a scoring function (s. next chapter), thus applying a selection-pressure towards poses with higher scores.[51]

While the ligand's conformational freedom is thus usually well accounted for in today's docking software, the protein is treated as rather rigid, which is a massive simplification. Already in the middle of the 20th century, it was noted that small molecules and proteins also bind to each other when their initial structures do not match well, leading to the "induced fit" model by Koshland which assumes that the structure of the receptor binding site is flexible and undergoes conformational changes to accommodate the ligand.[56] Later, based on the free energy landscape (FEL) theory of protein structure and dynamics,[57,58,59,60] the "conformational selection" model[61,62,63,64] was proposed which suggests that an ensemble of protein conformations exists in an equilibrium state, and that the ligand, by binding to the most suitable conformation, shifts this population. To address protein structural changes during docking, certain strategies have been developed. One approach which is typically applied is so-called "soft docking", which simply works by softening the interaction potential terms used in the respective scoring functions to allow a small degree of overlap between protein and ligand atoms to implicitly mimic small structural changes of the protein.[50] A step towards explicit treatment of protein flexibility is the use of rotamer libraries for amino acid sidechains, so that the side chain conformations can be varied for selected residues.[65] Another, rather intuitive way is to perform an ensemble docking into multiple available protein structures, either obtained from experiment or from MD simulations and homology modelling.[66,67]

Another factor often ignored in conventional docking is the treatment of binding site water molecules. As will be outlined in detail later, replacing or addressing water molecules can have huge effects on protein ligand binding. Consequently, studies showed that inclusion of selected water molecules during docking can lead to improved results.[68,69] The most straight-forward way to incorporate water is to treat all or selected experimentally determined water positions in a binding site as part of the protein. However, this might lead to bias in case that crystallographic waters are artefacts and can also limit the conformational space of possible ligand binding modes.[70,71] Besides, accurate scoring protocols must be used to accurately capture the respective contributions. A summary of methods concerning selection, prediction, and scoring w.r.t. water molecules in drug design will be given in chapter 2.4.

2.2.2.2.3 Scoring functions

As outlined earlier, the correct scoring of a ligand pose is essential for docking performance since it influences both the predicted binding mode of a single molecule and its ranking w.r.t. other compounds in a virtual screening experiment. Over the last decades, numerous scoring functions have been developed, which can be roughly differentiated into physics-based, empirical, knowledge-based, and machine learning-based.

Physics-based scoring functions, such as DOCK[72] and Goldscore,[73] attempt to approximate the binding (free) energy of a pose directly by utilising molecular mechanics (MM), solvation models, and/or quantum mechanical (QM) calculations. A classical approach are force field-based scoring functions which e.g. include terms for electrostatic and van der Waals interactions, bond stretching and bending, and torsions. To address solvation effects, additional implicit or explicit solvation models must be used.[74,75,76,77] In an attempt to address factors like polarisation, charge transfer, or the formation of covalent bonds, approaches based on QM or QM/MM have been developed; however, they remain computationally expensive.[78,79,80] Another problematic aspect in this context is to account for the change in conformational or configurational entropy of a given molecule upon binding.

Empirical scoring functions, too, estimate the binding free energy of a given protein-ligand complex structure, but in contrast to physics-based approaches, summation is performed over weighted simple, empirical energy terms $\Delta_{\mathrm{bind}}G_{\mathrm{LR},i}$, with the weighting-factors w_i being determined by regression on a training data set.

$$\Delta_{\mathrm{bind}}G_{\mathrm{LR}} = \sum_i w_i \Delta_{\mathrm{bind}}G_{\mathrm{LR},i} \tag{7}$$

Due to the simpler energy terms, empirical scoring functions are often more efficient; however, they rely on the quality of the respective training set.[50] A prominent example is ChemPLP,[81] the default scoring function used in GOLD, which was employed in this work. It includes terms for the steric complementarity of ligand and binding site (based on geometry and chemical properties), ligand torsion, angle- and distance-dependent hydrogen bonds and metal interactions, as well as terms describing clashes of atoms.

Knowledge-based scoring functions are based on the assumption that the frequency of a specific pair of atom types *i* and *j* at a specific distance *r* in a data set can be converted into a distance-dependent potential of mean force w_{ij} using the inverse Boltzmann relation. The binding free energy is then obtained by summation over all protein-ligand atom pairs in the complex according to:

$$\Delta_{\mathrm{bind}}G_{\mathrm{LR}} = \sum_{i=1}^{L}\sum_{j=1}^{R} w_{ij}(r) = -k_{\mathrm{B}}T\sum_{i=1}^{L}\sum_{j=1}^{R}\ln\left(\frac{\rho_{ij}(r)}{\rho_{ij}^{*}}\right) \tag{8}$$

Here, k_{B} is the Boltzmann constant, *T* the absolute temperature, $\rho_{ij}(r)$ the number density of the protein–ligand atom pair *ij* at distance *r*, and ρ_{ij} the pair density in a reference state.[82,83,84,85] Due to the pairwise evaluation, knowledge-based scoring functions are usually computationally efficient, and since they are based on a large number of complexes, they are rather robust and offer high accuracy especially in terms of binding pose prediction.[78] The main challenge w.r.t. knowledge-based scoring functions is the determination of a suitable reference state. This can for instance be achieved via randomisation of the atoms in the data set, as implemented e.g. in ASPScore[86] and DrugScore[87,88]. However, this technique neglects important factors like excluded volume;[85] therefore, approaches including correction terms or circumventing the reference state problem via a physics-based iterative method have been developed.[89,90,91,92,93,94]

The fourth category of scoring functions, based on ML techniques, is relatively new and is different from all the above-discussed "classic" approaches insofar that ML based-scoring functions do not have a defined, mathematical form with physically or statistically interpretable contribution terms. Hence, they are not expert- or theory-driven, but rather data-driven. The majority of ML algorithms perform as a "black box" which predicts a desired output, e.g. a binding free energy, via pattern recognition based on training on a provided data set. The advantage of this is that ML, if provided with enough high-quality data, has the potential to implicitly encode aspects of protein-ligand binding which are difficult to model explicitly.[95] However, the accuracy and robustness of the prediction heavily rely on the quality of the training data. Thus, a suitable training data set is a prerequisite for all ML approaches. From this data set, specific features are derived as input for the ML;[96] this can be e.g. fingerprints encoding structural features or interactions,[97] simple molecular descriptors,[98] SMILES,[99] molecule graphs,[100] individual energy terms from classical scoring,[101,102,103] or even 3D grids encoding molecular conformations.[104]

The most commonly applied techniques include support vector machines (SVMs), random forests (RFs), artificial neural networks (ANNs), and convolutional neural networks (CNNs).[105] SVMs map data into high-dimensional space via non-linear kernels and identify a hyperplane which leads to optimal separation of data points, e.g. into active and inactive.[106] The scoring function IDScore by Ding *et al.* is based on a modified support vector regression and utilises descriptors in form of categories related to protein-ligand interactions (e.g. electrostatic, hydrogen bond) to predict experimental binding affinities.[107]

As the name implies, a RF is a combination of a large number of decision trees and thus an ensemble learning method.[78] One of the most prominent ML-based scoring functions, RF-Score, uses simple geometric descriptors in form of protein-ligand atom pairwise counts in specific distances.[95] It has been mainly used for binding affinity prediction, and several versions using modified or additional descriptors have been published to improve accuracy.[97,108]

ANNs, in analogy to neural networks in biology, consist of individual neurons which are connected to the other neurons in a specific topology, e.g. several inter-connected layers including an input and output layer, and in many cases one or more hidden layers. Today, usually a NN is considered a “deep” NN if the number of hidden layers is more than three. Each neuron receives an input from the previous layer and applies a nonlinear function to it, with the resulting output being the input for the next layer. In the last layer, the output then corresponds to the desired prediction, e.g. a classification or a distinct value such as a binding free energy. During training, the error in the form of the difference between the network output and the “true” values is minimised via so-called backpropagation which adjusts the weights of the neuron connections.[109] The first NN-based scoring function, NNScore, takes a list of selected pairwise potentials describing for instance the electrostatic energy of different types of protein-ligand atoms pairs as input and was originally employed to classify docked compounds into binders and non-binders.[110] Since then, several improvements have been introduced, allowing for quantitative prediction of pK_d values rather than a mere classification.[111] Apart from retrospective validation, it has also been successfully applied in the development of ligands of several proteins relevant for medicinal chemistry.[112,113,114,115,116]

CNNs are a deep learning approach frequently applied in image recognition. Its hidden layer topology consists of convolutional layers, pooling layers and fully connected layers. Usually, the input to a CNN is a 2D or 3D matrix, e.g. colour channels of each pixel in an image or element information on a 3D grid of a protein-ligand complex. The convolutional layer converts this matrix input to abstracted feature maps which are then subsampled by the pooling layers to reduce redundancy. After several units of convolutional and pooling, the output is submitted to fully connected layers, in which every neuron is connected to each other neuron in the previous and next layer, to yield a final output in form of a prediction or classification.[96,117] One of the first CNN-based scoring functions is AtomNet which takes the 3D grid of a binding site as input, with each voxel containing information about respective structural features present in this voxel. It was evaluated on e.g. the DUD-E data set and showed good discrimination between active and inactive molecules.[118] The CNN-based scoring function K_{DEEP} was especially developed for predicting binding affinities and achieved an RMSE of around 1.3 pK units on the PDBbind core set 2016.[119] However, 3D CNNs are highly demanding w.r.t. GPU memory and storage. Therefore, also graph CNN-based approaches like the message passing NN (MPNN) GraphDelta have been developed which uses a molecule graph as input that encodes the 2D molecule structure as well as specific atom-based descriptors related to the interactions of the atoms in a protein-ligand complex.[100]

2.2.2.2.4 Assessment of scoring

When assessing the performance of a given scoring function from any of the four above-presented categories, several aspects have to be carefully considered: In general, there are four tasks based on which a scoring can be evaluated, namely docking power, ranking power, screening power, and prediction power.

Docking power describes the successful differentiation between “good” and “bad” poses of a single ligand, i.e. the capability of scoring the true binding mode above the other poses generated by the docking algorithm. Ranking power is the capability to correctly rank multiple ligands of a given target according to their experimental binding affinities. Screening power refers to the binary differentiation between active and inactive molecules and is of great importance in virtual screening, as presented e.g. in chapter 4.2 in this work. Screening power is usually evaluated by metrics like the enrichment-factor, which defines how many actives were enriched over inactives (“decoys”) at a certain percent of a ranked list compared to random classification, and the area under the curve of the receiver operating characteristic (ROC-AUC). The ROC curve is obtained by plotting the true positive rate (TPR) against the false positive rate (FPR) for each position i in the ranking that is obtained by sorting the active and decoy molecules according to their score. The TPR (also called “sensitivity” or “hit rate”) at a given position i in the ranking is defined according to:

$$\mathrm{TPR}_i = \frac{N_{\mathrm{active},i}}{N_{\mathrm{active,total}}}, \tag{9}$$

where $N_{\text{active},i}$ is the number of active compounds that have been successfully retrieved in the ranking up to position i, and $N_{\text{active,total}}$ the total number of actives in the data set. Hence, when iterating through the ranking positions, the value of TPR increases up to a value of 1.0 when all actives are found.

The FPR at a given position i in the ranking, consequently, is defined as:

$$\mathrm{FPR}_i = \frac{N_{\text{decoy},i}}{N_{\text{decoy,total}}}, \tag{10}$$

where $N_{\text{decoy},i}$ is the number of decoy compounds that have been retrieved in the ranking up to position i (and are thus false positives, since ideally all actives should be ranked before all decoys), and $N_{\text{decoy,total}}$ the total number of decoys in the data set. Hence, like the TPR, the value of FPR increases up to a value of 1.0.

For an ideal ranking that perfectly differentiates between actives and decoys, all actives would be ranked before all decoys, resulting in an TPR that grows to 1.0 while the FPR is still 0.0. This would result in the ROC "curve" being a horizontal line at TRP = 1.0, so that the corresponding AUC would have a value of 1.0. On the other hand, for a random classification of actives and decoys, TPR and FRP would grow alike, resulting in the ROC "curve" being the bisecting line, hence yielding an AUC of 0.5. Thus, ROC-AUC values near 1.0 denote a good differentiation between active and inactive molecules.

Prediction power (sometimes referred to as scoring power), on the other hand, defines the quantitative prediction of a binding free energy, often in form of a pK_d value, for a given protein-ligand complex structure from experiment or docking.[96]

Ideally, a scoring function would succeed equally well in all four tasks; however, different studies have shown that this is often not the case.[105,120] Several benchmark data sets have been established that can be employed for validation of new scoring approaches: The PDBbind database, which is regularly updated and was used in this work, is a selection of experimental complex structures from the PDB with available binding affinity values, with a high-quality core set comprising some hundred structures and a refined set with several thousand complexes.[121,122,123,124,125] It is thus especially suited for estimating the prediction power of scoring functions, but also docking power and ranking power. The Directory of Useful Decoys (DUD)[126] and DUD Enhanced (DUD-E),[127] as well as the Maximum Unbiased Validations (MUV) data sets[128] and DEKOIS data sets[129] provide collections of active and presumably inactive molecules for several proteins relevant in medicinal chemistry, thus allowing for the assessment of screening power.

The probably most prominent large-scale benchmark which is regularly carried out with the latest available scoring functions and current data sets is the Comparative Assessment of Scoring Functions

(CASF).[122,123] Evaluation of prediction power in CASF-2016 for instance revealed that the majority of classical, well-established scoring functions do not show satisfactory results in this task, with most Pearson correlation coefficients *R* being below 0.6, but perform much better for ranking and docking.[122] This might result from the fact that many classical scoring functions were not specifically developed to predict accurate, absolute binding affinities but rather to enrich good binders from poor binders, a task for which the prediction of absolute values is not necessarily essential. ML-based scoring functions, on the other hand, are primarily trained on experimental binding affinity data and are often not directly included in a docking software but rather applied as a post-docking rescoring. Interestingly, an assessment by Rognan *et al*. revealed that indeed the tested ML-based scoring functions outperformed classical scoring functions w.r.t. prediction power. However, they failed to discriminate between active and inactive molecules in the DUD-E data set and could not differentiate between native and other binding modes, thus failing w.r.t. both screening and docking power.[120] This study impressively highlights that care has to be taken concerning the robustness and applicability domain of newly developed approaches, especially for data-driven ML-methods.

2.2.3 MD simulations

All the approaches described so far rely on static 3D structures like an experimentally determined protein-ligand complex or an ensemble of these. However, in reality, molecular interactions are dynamic processes. Characterising these inherent dynamics can be of great advantage in SBDD, e.g. for the identification of so far untargeted cryptic binding pockets,[130,131] for optimising a ligand by assessing the overall stability of its interactions with the protein,[132] or even for studying the binding and unbinding process itself.[133,134,135] The most frequently applied approach for studying the dynamics of molecules computationally are MD simulations, in which atoms and their interactions are described via classic molecular mechanics (MM).

2.2.3.1 Theoretical background

In contrast to QM, in which the electronic structure of a system can in principle be completely described by a respective wave function, MM treats atoms as spheres which are connected via springs without differentiation between electrons and nuclei. Consequently, the energy of a given molecule can be stated as a function of the molecule's resistance w.r.t. distortions from the "natural" length and angle of its bonds. The entity of i) the mathematical expression of these geometry-based energy terms, and ii) the parameters used in them are defined as a force field.[136] Usually, they contain terms for energy contributions from bond stretching, angle bending, and torsions for covalent interactions and electrostatic and van der Waals interactions for non-bonded interactions. Over the decades, various force fields have been developed for different purposes; the most-widely used ones include CHARMM,[137,138] ff14SB,[139,140] GROMOS,[141] OPLS,[142] MMFF94s[143], and the general AMBER force field (GAFF).[144] In the latter, the energy *U* of a system is expressed in the following form:

$$U = \sum_{\text{bonds}} k_r (r - r_{\text{eq}})^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{v_n}{2}(1 + \cos(n\varphi - \gamma)) + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} + \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right] \quad (11)$$

Here, k_r, k_θ, and v_n are the force constants for bonds with the length *r*, bond angles with angle θ, and dihedrals with multiplicity *n*, angle ϕ and phase angle γ, respectively, with the subscript eq denoting equilibrium values; $q_{i,j}$, $A_{i,j}$, and $B_{i,j}$ are the partial charges and Lennard-Jones parameters of atoms *i* and *j* at distance R_{ij}; ε is the dieletric constant.[144] Compared to solving the Schrödinger equation, evaluation of the force field terms is far faster, thus allowing for the investigation of larger systems like protein-ligand complexes, which would be unfeasible with QM.

In MD simulations, the aim is to obtain a trajectory of the given system over time. The new position of an atom *i* after a time step Δt, $\mathbf{r}_i(t+\Delta t)$, can be approximated from the current position, $\mathbf{r}_i(t)$, using a Taylor series according to: [145]

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \frac{\partial \mathbf{r}_i(t)}{\partial t}\Delta t + \frac{1}{2}\frac{\partial^2 \mathbf{r}_i(t)}{\partial t^2}\Delta t^2 + O(\Delta t^3)\,, \quad (12)$$

with the second and third term being the atom’s velocity $\mathbf{v}_i$ and acceleration $\mathbf{a}_i$; *O* denotes the order of the error. The same can be applied for a step backwards in time. Addition of the resulting expression and Eq. (12) yields the Verlet algorithm, which is given by: [145,146]

$$\mathbf{r}_i(t + \Delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \Delta t) + \mathbf{a}_i(t)\Delta t^2 \quad (13)$$

and allows to determine the new position $\mathbf{r}_i(t+\Delta t)$ based on the current and last position of the atom and its acceleration. Following Newtonian mechanics, $\mathbf{a}_i$ can be obtained from the gradient of the potential *U* and its mass by:[145]

$$\mathbf{F}_i = -\partial U(\mathbf{r}_i, \ldots \mathbf{r}_N) / \partial \mathbf{r}_i = m_i \mathbf{a}_i\,, \quad (14)$$

with $\mathbf{F}_i$ denoting the force acting on atom *i*. Thus, the respective equations have to be solved for each atom in the system for each time step.

The length of this time step is limited by the vibration modes with the highest frequency of the system, like the hydrogen stretching, and accordingly is usually in the fs range.[147] Biologically interesting processes, like protein folding or ligand binding, exhibit time scales in the ns or µs range.[148] Hence, a huge number of time steps has to be simulated, resulting in high computational demand especially for very large systems containing e.g. membrane proteins. To reduce the computational cost, distance cutoffs can be introduced for the evaluation of the Lennard-Jones potential. Calculation of the Coulomb interactions can be improved via Ewald summation by separation into a short-range term, efficiently computable in real space using cutoffs, and a long-range term which is solved in Fourier space.[145]

Besides, to obtain physically relevant results, several other aspects have to be addressed in MD simulations: Since physiological processes take place in aqueous solution and not in vacuum, not only the molecule of interest has to be simulated, but also the surrounding solvent. This can be achieved via different explicit and implicit solvent models, which will be discussed more detailed in chapter 2.3. To avoid boundary effects due to the limited size of the simulation box, usually periodic boundary conditions are applied; here, the simulation box is treated as if it were surrounded by 26 identical copies of itself.[145]

When performing an MD simulation, an ensemble of configurations of the system is produced. If the configurational space of the system has been sampled sufficiently, this ensemble statistically represents the state of the given system and allows to derive specific properties as ensemble averages or integrals over the configurational space. If Newton's equations of motion are applied on a system of fixed volume *V* and atom number *N* without additional constraints, the total energy *E* of the system is constant, resulting in the microcanonical *NVE* ensemble. However, this does usually not well represent the experimental conditions of the studied systems of interest, so that methods have been developed to realise other thermodynamic ensembles. The most frequently employed ensembles are the canonical *NVT* ensemble with constant temperature *T*, and the isothermal-isobaric *NpT* ensemble with both temperature and pressure being held constant. To realise the *NVT* ensemble, the temperature has to be controlled by a thermostat algorithm which adds or removes energy e.g. by velocity scaling. Prominent approaches include the Nosé-Hoover thermostat,[149] the Berendsen thermostat,[150] or Langevin dynamics.[151] For approximating an *NpT* ensemble, an additional barostat, for instance the Berendsen[152] or Anderson barostat[153], is needed to control the pressure, e.g. by scaling of the box volume or interatomic distances.

2.2.3.2 Role of MD simulations in drug discovery

Unlike the approaches described in 2.2.2, MD simulation make it possible to explicitly capture structural flexibility and dynamics of the studied system, thus making it possible to study entropic effects and even the kinetics of protein-ligand binding.[154]

In the early beginnings of MD simulation applications for drug discovery, snapshots of obtained trajectories were used to provide input to classical SBDD methods: For instance, by extracting multiple binding side conformations from a trajectory, input structures for ensemble docking approaches were generated.[155] With growing computational power and thanks to graphical processor unit (GPU) architectures, MD simulations can now be run sufficiently long (up to µs or, in few examples, even ms range) to study the conformational space and dynamics of a protein-ligand complex.[154] For instance, it is nowadays often used for post-processing of docked protein-ligand complexes. Following the assumption that "bad" docking poses will result in unstable trajectories, the predicted binding modes can be validated.[154]

Yet, exploring the binding or unbinding process itself and the associated kinetics and free energy landscape is still a challenge since such processes often occur on a very long time scale.[156] To overcome this, enhanced sampling methods have been developed to make high-energy-states observable within the simulation. This includes, among others, free energy perturbation (FEP),[157] umbrella sampling,[158], replica exchange,[159] and steered MD.[160] With these approaches, it has become possible to study protein-ligand binding and the respective kinetics and energetics.

In drug discovery, FEP is also used to estimate the binding free energy difference between two related compounds. As the free energy is a state function, respective differences are independent of the path between two systems A and B. Hence, generally, for two different systems A and B, the difference in free energy can be defined as:[161]

$$G_{\mathrm{B}} - G_{\mathrm{A}} = \Delta G = -RT \ln \left\langle e^{-\Delta U/RT} \right\rangle_{\mathrm{A}}, \tag{15}$$

with U denoting a potential and the bracket term denoting an ensemble average over a system. For FEP, it is assumed that there is a parameter λ that can vary between 0 and 1, so that $U(\lambda)$ can be formulated as:[161]

$$U(\lambda) = \lambda U_{\mathrm{B}} + (1-\lambda) U_{\mathrm{A}}, \tag{16}$$

so that $U = U_{\mathrm{A}}$ for $\lambda = 0$ and $U = U_{\mathrm{B}}$ for $\lambda = 1$. Then, Eq. (15) can be generalized to:

$$G_{\mathrm{B}} - G_{\mathrm{A}} = \Delta G = \sum_{\lambda=0}^{1} -RT \ln \left\langle e^{-\Delta U'/RT} \right\rangle_{\lambda}, \tag{17}$$

with $\Delta U = U_{\lambda+\mathrm{d}\lambda} - U_{\lambda}$. Hence, the free energy calculation is performed by making small variations in λ.

Alternatively, in a thermodynamic integration (TI), a free energy difference between two pre-defined systems (characterized by $\lambda = 0$ or $\lambda = 1$) can be calculated by defining a thermodynamic path between these states and performing integration along this reversible path using λ as integration variable:

$$\Delta G = \int_0^1 \mathrm{d}\lambda \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \tag{18}$$

By evaluating the ensemble average of the derivative of the potential w.r.t. λ at various values of λ, numerical integration can be used to obtain the free energy difference.[161] Hence, such protocols can be employed for in-depth studies of identified lead candidates or for explaining SAR-trends. MD simulations have thus become a powerful tool that are nowadays employed in different stages of the SBDD process.

2.3 Solvation models

As outlined above, solvation effects are an important factor in all ligand-receptor binding events. The association constant K_b of a complex LR in aqueous solution as introduced in 2.1 can be calculated according to:[162]

$$\ln(K_\mathrm{b}) = \frac{\Delta_\mathrm{bind}G^{(\mathrm{g})} + \Delta_\mathrm{hyd}G_\mathrm{RL} - (\Delta_\mathrm{hyd}G_\mathrm{R} + \Delta_\mathrm{hyd}G_\mathrm{L})}{RT} \tag{19}$$

from the free energy of binding between ligand and receptor in gas phase, $\Delta_\mathrm{bind}G^{(\mathrm{g})}$, and the respective hydration free energies of ligand, receptor and complex, $\Delta_\mathrm{hyd}G_\mathrm{X}$. This formally corresponds to the thermodynamic cycle of first transferring R and L from solution to gas phase, forming the complex LR in gas phase and transferring the complex to solution (Figure 1).

$$\begin{tikzcd}
\mathrm{R}_{vac} + \mathrm{L}_{vac} \arrow[r, "\Delta_\mathrm{bind}G^{(\mathrm{g})}"] \arrow[d, "\substack{-\Delta_\mathrm{hyd}G_\mathrm{R} \\ -\Delta_\mathrm{hyd}G_\mathrm{L}}"'] & \mathrm{RL}_{vac} \arrow[d, "\Delta_\mathrm{hyd}G_\mathrm{RL}"] \\
\mathrm{R}_{aq} + \mathrm{L}_{aq} \arrow[r, "\Delta_\mathrm{bind}G^{(\mathrm{sol})}"'] & \mathrm{RL}_{aq}
\end{tikzcd}$$

Figure 1: Thermodynamic cycle for the calculation of $\Delta_{bind}G^{(sol)}$ by transferring R and L from solution to gas phase, forming the complex LR in gas phase and transferring the complex to solution.

Generally, Ben-Naim defined hydration as the transfer of a given solute molecule in a fixed position in an ideal gas phase to a fixed position in the aqueous phase at constant pressure and temperature. In the so-called Ben-Naim reference state, identical formal concentrations in solution and in gas phase are obtained. Hence, the hydration free energy $\Delta_\mathrm{hyd}G$ is defined as the respective change in the Gibbs free energy associated with this process; for arbitrary liquids other than aqueous solution it is called solvation free energy.[162,163] $\Delta_\mathrm{hyd}G$ can be separated into enthalpic and entropic contributions arising from the formation of bonds between solute and solvent and the accompanying increased ordering of the solvent molecules.

As Eq. (19) emphasises, hydration effects add a significant contribution to the free energy of binding of a complex in solution. However, in practice, this thermodynamic cycle can hardly be applied since the transfer to/from gasphase is usually associated with substantial structural, energetic and also entropic changes. Therefore, to account for hydration effects in simulations or energy calculations, different solvation models have been developed to approximate the effect of the solvent on the respective solute molecules. Generally, these approaches can be divided into explicit, implicit and hybrid models with varying focus onto accuracy and efficiency.

2.3.1 Explicit solvation models

The most straight-forward way to consider the solvent in any kind of simulation or calculation is to explicitly add solvent atoms to the system, e.g. by placing three-dimensional water molecules around a protein structure, and to treat their interactions exactly like those of the rest of the system. The obvious advantage is that interactions between the solvent and the solute are directly captured, thus allowing e.g. to identify hydrogen bonds between a conserved binding site water molecule and a protein residue. At the same time, this atomistic treatment leads to a massive increase of required calculations, making it costly w.r.t. computational power. Hence, atomistic treatment of the solvent in QM calculations is computationally very expensive. As a result, explicit solvation models are primarily used in MM approaches, like classical MD. Even here, idealistic models of the solvent (e.g. with fixed geometry and charges) are used to reduce the degrees of freedom. Frequently used water models are the TIPXP model (transferable intermolecular potential with X points),[164,165] and the SPC (simple point charge)[166] or SPC/E model (with an additional polarisation correction term).[166] An approach that allows to include electronic polarisation effects is *ab initio* MD (AIMD). Unlike in classical MD simulations, the forces needed for the generation of the trajectory are obtained from QM calculations: In each time step, electronic structure calculations are carried out using first-principles methods. Examples include the Born-Oppenheimer MD (BOMD) and the Car-Parrinello MD (CPMD).[167]

2.3.2 Implicit solvation models

An alternative to the accurate yet computationally expensive explicit treatment of individual solvent molecules is to approximate the overall effect that the solvent exerts on the solute without consideration of distinct solvent atoms. This type of methods is called implicit or continuum models.[168] Here, the solvent is described as a uniform medium with dielectric constant ε and a cavity for the embedded solute.[169]

Consequently, $\Delta_{\mathrm{hyd}}G$ can be decomposed according to:

$$\Delta_{\mathrm{hyd}}G = \Delta_{\mathrm{hyd}}G_{\mathrm{elec}} + \Delta_{\mathrm{hyd}}G_{\text{rep-disp}} + \Delta_{\mathrm{hyd}}G_{\mathrm{cav}} \tag{20}$$

into terms describing electrostatic interactions ($\Delta_{\mathrm{hyd}}G_{\mathrm{elec}}$), repulsion-dispersion interactions ($\Delta_{\mathrm{hyd}}G_{\text{rep-disp}}$), and the cavity formation ($\Delta_{\mathrm{hyd}}G_{\mathrm{cav}}$). Often, the two latter terms are combined to a non-polar contribution term $\Delta_{\mathrm{hyd}}G_{\mathrm{nonpolar}}$.[162] Usually, these models applied for a specific geometry, i.e. assuming a rigid body – otherwise, computationally more expensive ensemble averages are required (for instance by Molecular mechanics Poisson–Boltzmann surface area (MM-PBSA) or molecular mechanics generalized Born surface area (MM-GBSA) calculations).[170]

Electrostatic contributions to solvation can be described via continuum electrostatic theory by describing the solute as a cavity in a dielectric environment. The respective electrostatic potential **V**(**r**) for nonhomogeneous media (e.g solvent including ions) can be obtained by solving the Poisson-Boltzmann (PB) equation:

$$\nabla\left[\varepsilon(\mathbf{r})\nabla\mathbf{V}(\mathbf{r})\right] = -4\pi\rho^{f}(\mathbf{r}) - 4\pi\sum_{i} c_i^{\infty} z_i q \exp\left(-\beta z_i q \mathbf{V}(\mathbf{r})\right)\lambda(\mathbf{r}) \tag{21}$$

Here, ε is the dielectric constant of the solvent, $\rho^{\mathrm{f}}(\mathbf{r})$ the charge density of the solute including only molecular charges, c_i^{∞} the concentration of ion i at an infinite distance from the solute, z_i the valency of the ion, q the proton charge, β the reciprocal temperature, and $\lambda(\mathbf{r})$ the accessibility to ions at **r**.[171]

A computationally more efficient approach is the Generalised Born (GB) formalism which attempts to approximate the solution of the PB equation by describing the solute as a set of spheres with a specific dielectric constant. $\Delta_{\mathrm{hyd}}G_{\mathrm{elec}}$ can thus be approximated via:

$$\Delta_{\mathrm{hyd}}G_{\mathrm{elec}} = -\frac{1}{2}\left(1-\frac{1}{\varepsilon}\right)\sum_{i,j}\frac{q_i q_j}{f_{GB}}\,. \tag{22}$$

Here, q_i and q_j are are partial charges of atoms i and j, and f_{GB} is a function which interpolates an effective Born radius when the distance r_{ij} between i and j is short and the r_{ij} itself at large distances.[172]

$\Delta_{\mathrm{hyd}}G_{\mathrm{nonpolar}}$ is often estimated from the solute geometry, e.g. solvent accessible surface area (SASA) or solvent-exclusion volume, in combination with empirically determined proportionality parameters.[173,174,175] Commonly, the GB model estimating $\Delta_{\mathrm{hyd}}G_{\mathrm{elec}}$ is combined with a term estimating $\Delta_{\mathrm{hyd}}G_{\mathrm{nonpolar}}$ from SASA to approximate $\Delta_{\mathrm{hyd}}G$ in a so-called GB/SA approach.[176]

Generally, implicit solvent models allow for highly efficient calculations; however, effects like hydrogen bonds or reorientation of specific solvent molecules in the proximity of a solute cannot be directly captured since implicit models rather represent the bulk properties of the solvent.

2.3.3 Integral equation theory of molecular liquids

2.3.3.1 Classical density functional theory

Classical density functional theory (DFT) is a statistical mechanical theory to study the structure and thermodynamic properties of liquids that has many similarities to the well-known quantum DFT. In quantum DFT, the Hamiltonian of a multi-electron system in an external field (for instance due to the presence of the nuclei) is expressed as a unique functional of the electronic density. The electronic density of the ground state of the system minimises this functional and can thus be obtained via the variational principle.[177] The same principles also apply to classical systems, so that the presence of a solute in a liquid can be seen as an external potential $V_{\text{ext}}(\mathbf{r})$ that is exerted on the liquid.[178] Consequently, the Helmholtz energy F can be written as a functional of the particle density $\rho(\mathbf{r})$, with minimisation w.r.t. density yielding the equilibrium density $\rho_{\text{eq}}(\mathbf{r})$.

For a grand canonical ensemble, with constant T, V, and chemical potential μ, the grand potential Ω_V of a monoatomic liquid can be expressed as:[179]

$$\Omega_V[\rho(\mathbf{r})] = F[\rho(\mathbf{r})] + \int \rho(\mathbf{r})\{V_{\text{ext}}(\mathbf{r}) - \mu\}\mathrm{d}\mathbf{r}\,, \tag{23}$$

with the intrinsic Helmholtz free energy F. In case of the equilibrium density, Ω_V is minimised, i.e:

$$\left.\frac{\delta\Omega_V}{\delta\rho(\mathbf{r})}\right|_{\rho_{\text{eq}}} = 0 = \frac{\delta F}{\delta\rho(\mathbf{r})} + V_{\text{ext}}(\mathbf{r}) - \mu(\mathbf{r})\,, \tag{24}$$

with δ denotig the functional derivative, thus yielding:

$$\frac{\delta F}{\delta\rho(\mathbf{r})} \equiv \mu_{\text{in}}(\mathbf{r}) = \mu(\mathbf{r}) - V_{\text{ext}}(\mathbf{r})\,, \tag{25}$$

with $\mu_{\text{in}}(\mathbf{r})$ being defined as the so-called intrinsic potential which is the part of the chemical potential that does not depend on the external potential.

From Eq. (24), $\rho_{\text{eq}}(\mathbf{r})$ can thus be determined if an analytical expression of F is available. F can be separated into an ideal part F^{id} of a non-interacting, ideal fluid, and an excess part F^{ex} due to solvent-solvent interactions:[179]

$$F[\rho] = F^{\text{id}}[\rho] + F^{\text{ex}}[\rho] \tag{26}$$

The ideal part can be written as:[180]

$$F^{\mathrm{id}}\left[\rho\right]=-\beta^{-1}\int\rho\left(\mathbf{r}\right)\left(\ln\left[\Lambda^{3}\rho\left(\mathbf{r}\right)\right]-1\right)\mathrm{d}\mathbf{r}\,, \tag{27}$$

with Λ as the thermal wavelength; this expression is exact. The excess part, on the other hand, can be approximated using a Taylor series:[179]

$$F^{\mathrm{ex}}\left[\rho\left(\mathbf{r}\right)\right]=F^{\mathrm{ex}}\left[\rho_{\mathrm{eq}}\right]+\int\frac{\delta F^{\mathrm{ex}}\left[\rho_{\mathrm{eq}}\right]}{\delta\rho\left(\mathbf{r}\right)}\Delta\rho\left(\mathbf{r}\right)\mathrm{d}\mathbf{r}+\frac{1}{2}\int\frac{\delta^{2}F^{\mathrm{ex}}\left[\rho_{\mathrm{eq}}\right]}{\delta\rho\left(\mathbf{r}\right)\delta\rho\left(\mathbf{r}'\right)}\Delta\rho\left(\mathbf{r}\right)\Delta\rho\left(\mathbf{r}'\right)\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}'+\ldots \tag{28}$$

where $\Delta\rho = \rho$ - ρ_{eq}. The multi-body direct correlation functions $c^{(n)}$ can be generally defined as:[180]

$$c^{(n)}(\mathbf{r}_1,\mathbf{r}_2,\ldots\mathbf{r}_n)=-\frac{\delta\beta F^{\mathrm{ex}}}{\delta\rho(\mathbf{r}_1)\delta\rho(\mathbf{r}_2)\ldots\delta\rho(\mathbf{r}_n)} \tag{29}$$

Consequently, F^{ex} can be written as:

$$F^{\mathrm{ex}}\left[\rho\left(\mathbf{r}\right)\right]=F^{\mathrm{ex}}\left[\rho_{\mathrm{eq}}\right]-\beta^{-1}\int c^{(1)}\left(\mathbf{r}\right)\Delta\rho\left(\mathbf{r}\right)\mathrm{d}\mathbf{r}-\frac{1}{2}\beta^{-1}\int c^{(2)}\left(\mathbf{r},\mathbf{r}'\right)\Delta\rho\left(\mathbf{r}\right)\Delta\rho\left(\mathbf{r}'\right)\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}'+\ldots \tag{30}$$

The functional derivative of the ideal part yields:

$$\frac{\delta F^{\mathrm{id}}}{\delta\rho(\mathbf{r})}=-\beta^{-1}\ln\left[\Lambda^{3}\rho\left(\mathbf{r}\right)\right] \tag{31}$$

Together with the definition of the direct correlation function, the derivative of F is thus obtained as:

$$\mu_{in}\left(\mathbf{r}\right)=\frac{\delta F}{\delta\rho(\mathbf{r})}=\beta^{-1}\ln\left[\Lambda^{3}\rho\left(\mathbf{r}\right)\right]-\beta^{-1}c^{(1)}\left(\mathbf{r}\right). \tag{32}$$

The ideal gas chemical potential μ^{id} is defined as:[181]

$$\mu^{\mathrm{id}}=-\beta^{-1}\ln\left[\Lambda^{3}\rho(\mathbf{r})\right], \tag{33}$$

so that the term $\beta^{-1}c^{(1)}$ in Eq. (32) corresponds to the system's excess chemical potential in absence of an external field.

The functional derivative of μ_{in} w.r.t. $\rho(\mathbf{r})$ is linked to the inverse of the so-called density-density correlation function $H^{(\mathrm{n})}$ via:[179]

$$-\beta^{-1}\frac{\delta\mu_{in}(\mathbf{r})}{\delta\rho(\mathbf{r}')}=\frac{1}{\rho\left(\mathbf{r}\right)}\delta\left(\mathbf{r}-\mathbf{r}'\right)-c^{(2)}\left(\mathbf{r},\mathbf{r}'\right)\equiv H^{(2)-1}\left(\mathbf{r},\mathbf{r}'\right), \tag{34}$$

with the delta function δ. $H^{(2)}$ itself is defined as:

$$H^{(2)}\left(\mathbf{r},\mathbf{r}'\right)=-\beta^{-1}\frac{\delta\rho(\mathbf{r})}{\delta\mu_{\mathrm{in}}(\mathbf{r}')}=\rho\left(\mathbf{r}\right)\rho\left(\mathbf{r}'\right)h^{(2)}\left(\mathbf{r},\mathbf{r}'\right)+\rho\left(\mathbf{r}\right)\delta\left(\mathbf{r}-\mathbf{r}'\right), \tag{35}$$

where $h^{(2)}$ is the so called pair correlation function which is defined as:

$$h^{(2)}(\mathbf{r}_1,\mathbf{r}_2)=g^{(2)}(\mathbf{r}_1,\mathbf{r}_2)-1 \tag{36}$$

with the pair density distribution function $g^{(n)}(\mathbf{r})$ (also termed pair distribution function or pair correlation function):[162,179]

$$g^{(n)}\left(\mathbf{r}^{n}\right)=\frac{\rho^{(n)}(\mathbf{r}_{1},\ldots,\mathbf{r}_{n})}{\prod_{i=1}^{n}\rho(\mathbf{r}_{i})}. \tag{37}$$

Via combination of $H^{(2)}$ and $H^{(2)-1}$ and the functional definition of the delta function, the Ornstein-Zernike equation can be obtained, which is expressed for a uniform, isotropic liquid as:[162,179]

$$h\left(r_{12}\right)=c\left(r_{12}\right)+\rho\int h\left(r_{32}\right)c\left(r_{13}\right)\mathrm{d}r_{3}\ , \tag{38}$$

with r_{ij} denoting the distance between particles i and j. The OZ equation thus connects the total correlation function to the direct correlation function. The total correlation of two particles is formally separated into a direct correlation, i.e. the effect that particle 1 exerts directly on particle 2, and an indirect part that other particles, which are in turn influenced by the presence of particle 1, exert on particle 2. When trying to solve the OZ equation recursively by eliminating the h(r) in the integral, it becomes apparent that the right hand side of the equation is an infinite series of “chains” of different direct correlations:[162]

$$h\left(r_{12}\right)=c\left(r_{12}\right)+\rho\int c\left(r_{32}\right)c\left(r_{13}\right)\mathrm{d}r_{3}+\rho^{2}\int c\left(r_{34}\right)c\left(r_{13}\right)\mathrm{d}r_{3}\,\mathrm{d}r_{4}+\ldots \tag{39}$$

Since the OZ equation contains two unknown functions, $h(r)$ and $c(r)$, a second equation is needed to solve it, which is called “closure relation”. This closure relation can be expressed as:[179,182,183]

$$h(r)+1=g(r)=\exp(-\beta u(r)+h(r)-c(r)+B[t(r)]) \tag{40}$$

Here, $u(r)$ is the pair interaction potential between the particles, and $B[t(r)]$ a so-called bridge function which is a functional of the indirect correlation function $t(r) \equiv h(\mathrm{r}) - c(r)$. However, the correct expression of $B[t(r)]$ is unknown, so that finding a suitable approximation is a main challenge in this field. The simplest approximation is to set $B[t(r)] = 0$; this corresponds to the so-called hypernetted chain (HNC) closure.[183]

The equations presented above describe monoatomic, isotropic liquids. To extend the theory to non-spherical molecules, the dependence of the correlation from the respective orientation has to be considered. This is expressed in the so-called Molecular Ornstein Zernike (MOZ) equation:[162]

$$h(\mathbf{r}_{12},\Theta_{1},\Theta_{2})=c(\mathbf{r}_{12},\Theta_{1},\Theta_{2})+\frac{\rho}{\mathbb{Z}}\int c(\mathbf{r}_{13},\Theta_{1},\Theta_{3})h(\mathbf{r}_{32},\Theta_{3},\Theta_{2})\mathrm{d}\mathbf{r}_{3}\mathrm{d}\Theta_{3} \tag{41}$$

Here, Θ_1 und Θ_2 are sets of the Euler angles denoting the orientation of the two molecules 1 and 2 to each other, and $\mathbb{Z}$ is 4π for linear molecules or $8\pi^2$ for non-linear molecules.

When regarding a solute molecule in pure solvent at infinite dilution, it is possible to express the correlation via three independent equations for solvent-solvent, solute-solvent, and solute-solute correlation functions.[183]

2.3.3.2 Reference interaction site model

Solving the MOZ in Eq. (41) in principle allows for the direct calculation of solvent structure; however, in practice the high dimensionality leads to several problems, so that methods have been developed to reduce the dimensionality of the MOZ. They are based on the work of Chandler and Andersen[184] and are termed reference interaction site models (RISM).[162]

In the RISM approach, solute and solvent molecules are considered as a set of spherically symmetric interactions sites (which can correspond e.g. to the molecule's individual atoms) that can be described by a set of site-site correlation functions, so that the 6D MOZ is reduced to several 1D equations. Thus, there are three types of correlation functions which depend only on the radial distance r between these sites: the intramolecular correlation functions $\omega(r)$, the direct correlation functions $c(r)$, and the total correlation functions $h(r)$.[162,185]

The intramolecular correlation $\omega(r)$ for two sites α and α' of one molecule at distance $r_{\alpha\alpha'}$, it is given by:[162]

$$\omega_{\alpha\alpha'}(r) = \frac{\delta(r - r_{\alpha\alpha'})}{4\pi r_{\alpha\alpha'}^2} . \tag{42}$$

The intermolecular direct correlation function $c_{\alpha\gamma}$(r) of two sites α and γ is approximated via the sum over the individual site-site correlation functions:[183]

$$c_{\alpha\gamma}(r) = \sum_{\alpha}\sum_{\gamma} c_{\alpha\gamma}(r_{\alpha\gamma}) \tag{43}$$

Thus, the total correlation functions can be obtained from the intramolecular and the direct correlation functions via a set of 1D equations. This resulting RISM equation can be expressed in matrix form as:[185]

$$\mathbf{h} = \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\omega} + \boldsymbol{\rho}\boldsymbol{\omega} * \mathbf{c} * \mathbf{h} , \tag{44}$$

where * denotes a matrix convolution. In general, the convolution $f*g$ of two functions f, g: $\mathbb{R}^n \rightarrow \mathbb{C}$ is defined as:

$$(f * g)(x) := \int_{\mathbb{R}^n} f(\tau) g(x - \tau) \mathrm{d}\tau . \tag{45}$$

For discrete functions f, g: $D \rightarrow \mathbb{C}$, the integral is replaced by a summation:

$$(f * g)(x) := \sum_{\tau \in D} f(\tau) g(x - \tau) \tag{46}$$

When considering not only the pure solvent, but mixtures of solvent with a solute at infinite dilution, the respective RISM equation is given by:

$$\boldsymbol{\rho}\mathbf{h}\boldsymbol{\rho} = \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\omega} + \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\rho}\mathbf{h}\boldsymbol{\rho} \tag{47}$$

The respective matrices can be reorganised into blocks containing only either the solvent-solvent (*vv*), solute-solvent (*uv*) or solute-solute (*uu*) correlation functions.

$$\mathbf{h}^{vv} = \boldsymbol{\omega}^{v} * \mathbf{c}^{vv} * \boldsymbol{\omega}^{v} + \boldsymbol{\omega}^{v} * \mathbf{c}^{vv} * \boldsymbol{\rho}^{v}\mathbf{h}^{vv} \tag{48}$$

$$\mathbf{h}^{uv} = \boldsymbol{\omega}^{u} * \mathbf{c}^{uv} * \boldsymbol{\omega}^{v} + \boldsymbol{\omega}^{v} * \mathbf{c}^{uv} * \boldsymbol{\rho}^{v}\mathbf{h}^{vv} \tag{49}$$

$$\mathbf{h}^{uu} = \boldsymbol{\omega}^{u} * \mathbf{c}^{uu} * \boldsymbol{\omega}^{u} + \boldsymbol{\omega}^{u} * \mathbf{c}^{uv} * \boldsymbol{\rho}^{v}\mathbf{h}^{uv} \tag{50}$$

As can be seen, the resulting equations are hierarchical, so that $\mathbf{h}^{vv}$ can be used for solving $\mathbf{h}^{uv}$, which can in turn be employed for solving $\mathbf{h}^{uu}$. The solvent-solute equation can be rewritten as:

$$\mathbf{h}^{uv} = \boldsymbol{\omega}^{u} * \mathbf{c}^{uv} * (\boldsymbol{\rho}^{v})^{-1}(\boldsymbol{\rho}^{v}\boldsymbol{\omega}^{v} + \boldsymbol{\rho}^{v}\mathbf{h}^{vv}\boldsymbol{\rho}^{v}) \tag{51}$$

to separate the solvent-solute related matrices from those only related to solvent-solvent correlations. The term in brackets is defined as the solvent-susceptibility function $\boldsymbol{\chi}$:

$$\boldsymbol{\chi} = \boldsymbol{\rho}^{v}\boldsymbol{\omega}^{v} + \boldsymbol{\rho}^{v}\mathbf{h}^{vv}\boldsymbol{\rho}^{v} \tag{52}$$

It has to computed only once for a given solvent and can then conveniently be used for any *uv* calculation.

The above outlined 1D RISM equations can be converted to three dimensions by replacing the spatial distribution functions by radial distribution functions, so that the molecular orientation of the solvent around the infinitely diluted solute is averaged out. For the *uv* case, this results in:[186,187]

$$\rho_{\gamma}^{0} h_{\gamma}(\mathbf{r}) = \sum_{\gamma'} c_{\gamma'}(\mathbf{r}) * \chi_{\gamma\gamma'}(|\mathbf{r}|), \tag{53}$$

where $\chi_{\gamma\gamma'}$ is the pure solvent susceptibility, which can be precomputed by 1D RISM, and ρ_{γ}^{0} the bulk density. Consequently, the respective closure relations (one per solvent site) are given by:

$$h_{\gamma}(\mathbf{r}) = \exp\left(-\beta u_{\gamma}(\mathbf{r}) + h_{\gamma}(\mathbf{r}) - c_{\gamma}(\mathbf{r}) + B_{\gamma}(\mathbf{r})\right) - 1 \tag{54}$$

A highly relevant property which can be calculated by RISM is the excess chemical potential, μ^{ex}. Formally, it can be derived from the coupling parameter integral:[188]

$$\mu^{\mathrm{ex}} = \int_{0}^{1} \mathrm{d}\lambda \sum_{\gamma} \int \mathrm{d}\mathbf{r}\rho_{\gamma}(\mathbf{r},\lambda) \frac{\mathrm{d}u_{\gamma}(\mathbf{r},\lambda)}{\mathrm{d}\lambda} \quad , \tag{55}$$

with $\rho_{\gamma}(\mathbf{r},\lambda) = \rho_{\gamma}g_{\gamma}(\mathbf{r},\lambda)$, and $\mathrm{d}u/\mathrm{d}\lambda$ denoting the partial derivative of the interaction potential u_{γ} between solute and solvent w.r.t. the coupling parameter λ. Solving this equation directly would be computationally costly since it would require solving it for every λ step. However, using RISM and respective closure relation approximations, the expression for μ_{ex} is an integral over an exact differential which can hence be solved analytically. For the HNC-closure, this yields:[162]

$$\mu_{\text{HNC}}^{\text{ex}} = \int \mathrm{d}\mathbf{r} \sum_{\gamma} \frac{\rho_{\gamma}^{0}}{\beta} \left(\frac{1}{2} h_{\gamma}^{2}(\mathbf{r}) - c_{\gamma}(\mathbf{r}) - \frac{1}{2} h_{\gamma}(\mathbf{r}) c_{\gamma}(\mathbf{r}) \right) \tag{56}$$

As outlined above, the HNC closure ignores the bridge function by approximating it as 0, which is a suitable approximation for large distances but can lead to numeric instabilities. In this work, the "partial series expansion" of order *n* (PSE-*n*) developed by Kast and Kloss was applied for *uv* RISM calculations:[188,189]

$$h_{\gamma}(\mathbf{r}) = \begin{cases} \exp[t_{\gamma}^{*}(\mathbf{r})] - 1 & \Leftrightarrow t_{\gamma}^{*}(\mathbf{r}) \le 0 \\ \sum\limits_{n} [t_{\gamma}^{*}(\mathbf{r})]^{n} / n! - 1 \Leftrightarrow t_{\gamma}^{*}(\mathbf{r}) > 0 \end{cases}, \tag{57}$$

with $t_{\gamma}^{*}(\mathbf{r}) = h_{\gamma}(\mathbf{r}) - c_{\gamma}(\mathbf{r}) - \beta u_{\gamma}(\mathbf{r})$. The PSE-*n* closure satisfies the HNC closure approximation for $n \to \infty$. The interaction potential $u_{\gamma}(\mathbf{r})$ between a solvent site γ and the solute is given by:[162]

$$u_{\gamma}(\mathbf{r}) = \sum_{\alpha} u_{\alpha\gamma}\left(\left|\mathbf{r}_{\alpha} - \mathbf{r}\right|\right), \tag{58}$$

with:

$$u_{\alpha\gamma}\left(\left|\mathbf{r}_{\alpha} - \mathbf{r}_{\gamma}\right|\right) = u_{\alpha\gamma}^{\text{LJ}} + u_{\alpha\gamma}^{\text{elec}} = \sum_{\alpha\gamma} 4\varepsilon_{\alpha\gamma} \left(\left(\frac{\sigma_{\alpha\gamma}}{\left|\mathbf{r}_{\alpha} - \mathbf{r}_{\gamma}\right|} \right)^{12} - \left(\frac{\sigma_{\alpha\gamma}}{\left|\mathbf{r}_{\alpha} - \mathbf{r}_{\gamma}\right|} \right)^{6} \right) + \sum_{\alpha\gamma} \frac{q_{\alpha} q_{\gamma}}{4\pi\varepsilon_{0} \left|\mathbf{r}_{\alpha} - \mathbf{r}_{\gamma}\right|}. \tag{59}$$

Here, $\sigma_{\alpha\gamma}$ and $\varepsilon_{\alpha\gamma}$ are the mixed Lennard-Jones parameters and q_{α} and q_{γ} the partial charges of the respective solute and solvent sites. Usually, Ewald summation is used for enhanced performance by separating the electrostatic potential into a short-range part, which can be solved in real-space, and a long-range part, which can be solved in reciprocal space after Fourier transformation.

For the PSE-*n* closure, the expression for μ^{ex} is given by:[188,190]

$$\mu_{\text{PSE-}n}^{\text{ex}} = \int \mathrm{d}\mathbf{r} \sum_{\gamma} \frac{\rho_{\gamma}^{0}}{\beta} \left[\frac{1}{2} h_{\gamma}^{2}(\mathbf{r}) - c_{\gamma}(\mathbf{r}) - \frac{1}{2} h_{\gamma}(\mathbf{r}) c_{\gamma}(\mathbf{r}) - \Theta(h_{\gamma}(\mathbf{r})) \frac{(t_{\gamma}^{*}(\mathbf{r}))^{n+1}}{(n+1)!} \right] \tag{60}$$

with the Heaviside function Θ.

Within the applied approximations (e.g. neglecting reorganisation and polarization of the solute upon solvation), the μ^{ex} in water corresponds to the hydration free energy $\Delta_{\text{hyd}}G$. The ensemble-independent $\Delta_{\text{hyd}}G$[191] can thus be defined as an integral over a free energy density $\rho_{\text{G}}(\mathbf{r})$ according to:[192]

$$\mu_{\text{PSE-}n}^{\text{ex}} = \Delta_{\text{hyd}}G \equiv \int \mathrm{d}\mathbf{r} \rho_{G}(\mathbf{r}). \tag{61}$$

In this work, this is of special interest since it allows to determine the contribution of a spatial region to the total hydration free energy. For instance, when considering a protein binding site in water, the local information about the solvent density, provided by $g_\gamma(\mathbf{r}) = h_\gamma(\mathbf{r}) + 1$, permits to draw conclusions about the position of specific hydration sites. Integration of $\rho_\mathrm{G}(\mathbf{r})$ over the spatial region corresponding to one water molecule then yields the contribution of this specific water molecule w the free energy of hydration of the protein, $\Delta_\mathrm{hyd}G_\mathrm{P}$. This procedure allows to predict unstable water molecules within a binding site and can be used to get a detailed picture of the thermodynamic signature of *apo* (but also *holo*) protein binding sites which is highly relevant for SBDD.

3D RISM *uu* calculations treating two solutes 1 and 2 in solvent require the individual *uv* results of both reactants in the respective solvent, namely $c^{uv}_{1,\gamma}$ ($\mathbf{r}$; $\mathbf{R}_1$, Ω_1) and $\rho_\gamma h^{uv}_{2,\gamma}$ ($\mathbf{r}$; $\mathbf{R}_2$, Ω_2), which depend on their positions and orientations ($\mathbf{R}_i$, Ω_i) in the 3D grid ($\mathbf{r}$). The respective expression for h^{uu} can then be formulated using the distance vectors and Euler angles $\mathbf{R}_{12}$ and Ω_{12} of a "super-molecule" consisting of solute 1 and solute 2 in fixed relative geometry:

$$h^{uu}(\mathbf{R}_{12},\Omega_{12}) = c^{uu}(\mathbf{R}_{12},\Omega_{12}) + \sum_\gamma c^{uv}_{1,\gamma} * \rho_\gamma h^{uv}_{2,\gamma}(\mathbf{R}_{12},\Omega_{12})\,. \qquad (62)$$

An implementation of the 3D RISM *uu* formalism was developed in the Kast group by F. Mrugalla and was used in the past for designing molecular complexes using free-energy derivatives: As shown in the respective work, the solute-solute equation of RISM allows to compute derivatives of the potential of mean force (PMF) w.r.t. potential parameters, thus allowing to define an optimisation direction in the chemical space towards and optimised binding of the respective molecular partners.[259] In this work, the *uu* formalism is of special interest for the second part of studies, since it allows for a detailed characterisation of binding sites by determining the local distribution of pharmacophoric probes: Based on the 3D *uv* calculation of a protein and a 1D RISM *uv* calculation of a simple, spherical probe that mimics a ligand functional group (for instance a positively charged nitrogen), a 3D RISM *uu* calculation can be performed to obtain the pair distribution function of the respective probe within the binding site. By doing this with different probes representing distinct pharmacophoric features, this yields a detailed profile of the binding site w.r.t. physicochemical properties of suitable ligand groups.

Besides, RISM theory can also be combined with quantum chemical calculations. In the Kast group, the so-called embedded cluster RISM (EC-RISM) was implemented and successfully applied for quantitative prediction of pK_a, log$P$, and log$D$ values.[193,194,198] While 3D RISM gives access to the excess chemical potential in solution, it does not yield the intramolecular energy of the molecule polarised by the solvent since this is not possible with fixed charge force fields. Hence, polarisable forcefields or quantum chemical calculations – as in EC-RISM – are necessary to obtain this intramolecular energy. Within the EC-RISM framework, an iterative cycle is carried out: First, an electrostatic potential is determined from the vacuum wave function of the fixed solute. As a second step, the polarised solvent distribution based on this electrostatic potential is determined. Afterwards, the solute is embedded in a set of point charges that represent the point charges, thus yielding a new wave function and electrostatic potential. The second and third step are repeated until self-consistency of electronic and solvent structure is reached.

All in all, the RISM theory thus offers a powerful framework to obtain the equilibrium structure, and hence thermodynamic properties, of systems in solution at significantly less computational effort than for instance explicit solvent MD simulations (using a rigid body approximation). In MD, an equilibrium structure has to be calculated as averages over a large number of individual configurations, requiring sufficient sampling and thus long simulation times. With RISM, in contrast, this equilibrium solvent structure is obtained based on a single configuration of a system while still reflecting the atomistic characteristics of the solvent, which is not the case for continuum methods.[162]

2.3.3.3 Empirical corrections for 3D RISM

When calculating absolute values of the excess chemical potential using 3D RISM, a well-known artifact has to be accounted for which usually leads to values which are too high.[162,195,196,197] A reason for this is the overestimation of the energy required to form a cavity in the solvent.[162] This error however shows an almost linear correlation with the solute's partial molar volume V_m, a thermodynamic quantity which describes the variation of a solution's volume upon addition of the solute and which can be readily obtained by 3D RISM calculations.[162] Hence, a rather simple correction term can be defined to account for this overestimation.

In addition, in case of an ionic solute, another aspect has to be considered: Within 3D RISM, the solvent is infinite and hence does not have a surface, so that surface polarisation is neglected. To account for this, an additional correction term based on the solute's charge q can be defined.

Accordingly, similar to work by others,[162,196] a respective correction for the absolute value of μ^ex was formulated in the author's working group that takes both aspects into account:[195,198]

$$\mu^{\mathrm{ex,corr}} = \mu^{\mathrm{ex}} + c_V V_m + c_q q \tag{63}$$

Here, μ^ex is the uncorrected excess chemical potential of a given solute in a solvent as obtained by 3D RISM, V_m the solute's infinite dilution partial molar volume as obtained by 3D RISM, and q the solute's charge. The correction parameters c_V and c_q have to be obtained via regression on available experimental data for a given solvent (for instance experimental solvation free energies).

Although the present work is not concerned with the prediction of absolute hydration free energies but rather with the analysis of respective local contributions of specific spatial regions, it might be of interest to investigate if and how the described correction affects the underlying $\rho_G(\mathbf{r})$ field as described in Eq. (60) and (61). This can be achieved by performing the V_m-based correction as defined in Eq. (63) not on absolute values, but on grid level. This approach for a local PMV correction on grid level is presented in the following and was applied on a selected protein structure in a proof-of-concept study within this work.

According to site-site Kirkwood-Buff theory, the absolute V_m of a solute in a given solvent at infinite dilution can be obtained from the site-site direct correlation function $c_{\alpha\gamma}$ as:[199]

$$V_m = \beta^{-1}\kappa\left(1-\rho\sum_{\alpha,\gamma}\int c_{\alpha\gamma}(\mathbf{r})4\pi r^2 \mathrm{d}\mathbf{r}\right) \tag{64}$$

Here, κ denotes the pure solvent's isothermal compressibility (which can be obtained from 1D RISM), ρ the pure solvent's density, and $c_{\alpha\gamma}$ the site-site direct correlation function of solvent and solute sites α and γ.[198] Consequently, within the 3D RISM framework, a local $V_m(\mathbf{r}_i)$ can be evaluated at each volume element $\mathbf{r}_i$ of the grid $\mathbf{r}$ according to:

$$V_m(\mathbf{r}_i) = \frac{\beta^{-1}\kappa}{N_\mathbf{r}} - \beta^{-1}\kappa\rho V(\mathbf{r}_i)\sum_{\alpha,\gamma} c_{\alpha\gamma}(\mathbf{r}_i), \tag{65}$$

with $N_\mathbf{r}$ denoting the total number of volume elements in the grid and $V(\mathbf{r}_i)$ the volume of one volume element (please note that $\mathbf{r}_i$ refers to a specific volume element in this case and not an atom).

Hence, at each volume element, a local V_m- and q-corrected μ^{ex} contribution, $\mu^{\text{ex,corr}}(\mathbf{r}_i)$, can be obtained by applying the correction defined in Eq. (66):

$$\mu^{\text{ex,corr}}(\mathbf{r}_i) = \mu^{\text{ex}}(\mathbf{r}_i) + c_V V_m(\mathbf{r}_i) + c_q \frac{q}{N_\mathbf{r}}, \tag{66}$$

with the side condition that:

$$\sum_i \mu^{\text{ex,corr}}(\mathbf{r}_i) = \mu^{\text{ex,corr}} \cap \sum_i V(\mathbf{r}_i) = V, \tag{67}$$

with V being the total volume of the grid.

By summing up the $\mu^{\text{ex}}(\mathbf{r}_i)$ or $\mu^{\text{ex,corr}}(\mathbf{r}_i)$ contributions of volume elements around a given water position $\mathbf{r}_w$, an estimate of the respective contribution of that water molecule w to the total free energy of hydration of the protein, $\Delta_{\text{hyd}}G_{\text{P},w}$, is obtained. Comparison of the resulting $\Delta_{\text{hyd}}G_{\text{P},w}$ values for the original and corrected $\rho_G(\mathbf{r})$ field allows to estimate if the correction affects the local free energy distribution (e.g. if it leads to a switch of an unfavourable $\Delta_{\text{hyd}}G_\text{P}$ contribution to a favourable one or vice versa). However, as will be shown later in this work, it was found that the local V_m- and q-correction does not lead to relevant inversions in the free energy distribution.

2.4 Water in SBDD

In drug design, the relevance of water molecules as a “third party”[200] in protein-ligand binding has increasingly been recognised. Several case studies are known where displacement of structural water molecules plays an important role, for instance the cyclic urea series of HIV-1 protease inhibitors[201] or the 5-cyanopyrimidine derivatives of p38α MAP kinase inhibitors.[202]

Whether the effect of replacing or targeting a hydration site on the free energy of binding, $\Delta_{\text{bind}}G_{\text{PL}}$, is favourable or unfavourable depends on different factors:[192] If a water molecule with an unfavourable contribution to the hydration free energy of the protein, $\Delta_{\text{hyd}}G_{\text{P}}$, is displaced, this will be beneficial for $\Delta_{\text{bind}}G_{\text{PL}}$. If a ligand atom displaces a water molecule with a favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contribution, on the other hand, this loss has to be compensated by favourable enthalpy contributions $\Delta_{\text{bind}}H_{\text{PL}}$, i.e. favourable interactions between the ligand and the binding site residues. Water molecules that are not displaced can still contribute to $\Delta_{\text{bind}}G_{\text{PL}}$ due to interactions with the ligand.

Analyses on experimental protein-ligand complex structures have shown that water can have a stabilising effect by bridging interactions between the ligand and the binding site residues.[203,204] Besides, the presence of water networks with many H-bonds around the ligand was found to correlate with improved binding affinity.[205,206,207] Thus, for rationally modulating ligand affinity, it is highly desirable to have information about the positions and thermodynamic properties of hydration sites in the respective protein binding site. Terminologies like “happy” and “unhappy” or “cold” and “hot” waters have been coined to intuitively capture the relevance of water thermodynamics for drug design purposes.[208] A very prominent example is WaterMap by Schrödinger which will be in detailed explained below.[209,210,211]

The first step towards including water molecules in rational drug design is the classification of experimental water positions, e.g. from X-ray crystallography, into conserved and displaceable ones. Two of the first programs for such a discrimination include Consolv[212] and WaterScore[213] which are based on analysing temperature B-factors and the protein environment of experimental structures. PyWATER,[214] a PyMOL plugin, identifies conserved water molecules in a protein structure based on superposition with structures of the same protein family. A similar approach was already pursued 1998 by WatCH (Waters Clustered Hierarchically).[215] WaterRank[216] provides a conservation/nonconservation classification of water molecules based on HINT (Hydropathic INTeractions)[217] and a geometric descriptor.

When no experimental water positions are available, they have to be predicted by suitable methods. Knowledge based approaches for prediction of water positions include AQUARIUS,[218] AcquaAlta,[219] the tetrahedron-water-cluster model,[220] and WarPP[221], which was validated on approx. 1500 protein structures. WATGEN[222] was especially developed for describing water networks at protein protein interfaces. WaterDock,[76] a docking based approach, utilises AutoDock Vina for hydration site prediction and a combination of data-mining, heuristic and machine learning techniques for their classification into displaced and conserved sites.

The so far described methods mainly focus on a binary conserved/displaceable classification. However, as will be shown later in this work, the question if a water can be replaced or not might actually be influenced by the properties of the replacing group. Therefore, it is beneficial to obtain quantitative information about the water molecules' thermodynamic properties rather than a mere classification. The methods suitable to obtain such quantitative information can be roughly categorised into simulation or grid-based methods.

In general, simulation-based methods generate a huge number of possible configurations of the respective system including explicit solvent atoms, which are then subjected to statistical analysis. Hence, simulation-based approaches require significant computational cost and might suffer from insufficient sampling but allow for a detailed description of individual interactions and water networks.[223,224,225]

The probably most popular and widely used method for predicting hydration sites and their thermodynamic properties is WaterMap[209,210,211] by Schrödinger which combines MD simulation with Inhomogeneous Solvation Theory (IST):[226,227] Usually, a short (approx. 2 ns) MD simulation is performed with the protein held rigid. Based on the resulting trajectory, populations are determined via a cluster analysis. The thermodynamic properties of each hydration site are then calculated via IST by approximating the respective average interaction energy contribution and the entropic penalty resulting from the decrease in the degrees of freedom. WaterMap was successfully applied in several case studies in medicinal chemistry.[210,211,228,229] A similar approach is the combination of MD with grid-based IST (GIST) as developed by Gilson *et al.*,[230,231] where the energy and entropy contributions are discretised onto a 3D grid. This approach was integrated into the scoring function AutoDock4 to directly use it in docking.[77] Besides, it was employed to predict hydration free energies,[232] and to identify spatial regions within a system with specific enthalpic or entropic properties.[233] Other MD-based approaches include STOW,[234] SPAM,[235] WATsite,[236] the Two-Phase Thermodynamic (2PT) Model,[237] and the "cell-theory" ansatz by Henchman.[238]

Another set of simulation-based methods relies on employing Monte Carlo (MC) simulations, e.g. grand canonical MC (GCMC),[239,240,241] MC reference state (MCRS),[242] and JAWS.[243] Barillari *et al*. employed MC using replica exchange thermodynamic integration and the double decoupling approach for calculating binding free energies of water molecules and showed that conserved water molecules are generally more tightly bound.[244] In these approaches, usually no larger conformational changes are allowed, so that proteins are treated as rather rigid.

In contrast to simulation-based methods, grid-based approaches do not generate multiple configurations of the system of interest (i.e. a protein in water) but try to generate the equilibrium configuration. They are usually much faster than simulation-based methods but can lack the atomistic detail (depending on the used theory). One prominent example is WaterFLAP[43] by Molecular Discovery. It utilises the GRID molecular interaction fields to identify the position and interaction energy of a respective spherical water probe. SZMAP[245,246] by OpenEye relies on the calculation of a Poisson-Boltzmann potential on a grid, followed by placement and energy evaluation of a water probe which in contrast to the WaterFLAP probe can have different orientations. 3D RISM,[184,186,187,247] which is employed in this work and is in detail described in 2.3.3, is another grid-based method. It allows to estimate the equilibrium density of water (or any other solvent) around a solute and provides an analytical expression of the free energy of hydration, which can be evaluated w.r.t. local contributions of specific hydration sites.[248,249,250,251] Thus, it offers more atomistic detail than continuum models at significantly less computational cost than simulation-based approaches.

2.5 Aims and approaches of this work

In the last sections, an overview of SBDD methods was presented, with a focus on the relevance and treatment of solvation effects. The investigation of these solvent effects, and the deduction of rules to exploit knowledge about the local thermodynamic properties of protein hydration sites for drug design purposes, is the main objective of this work. The method of choice to achieve this is the afore-presented 3D RISM theory. As explained in 2.3.3., it does not only allow to predict the solvent distribution within a binding site, but also to calculate local thermodynamic properties of specific hydration sites.

The first part of this work therefore focuses on a large-scale analysis of the thermodynamic signatures of protein hydration sites and their correlation with ligand features. Water replacement rules for use in ligand design and optimisation are derived and exemplified on the basis of matched molecular pairs (MMPs).

The second part of this work expands the concept of 3D RISM-derived thermodynamic binding site characterisation to virtual probe sites that mimic specific functional groups whose distribution within a binding site can be calculated by the afore-presented 3D RISM solute-solute approach implemented in the Kast group. An advanced framework combining RISM-based binding site characterisation with automatised library preparation, docking, and scoring is established that allows to "convert" the probe densities to a selection of promising fragments or small molecules, thus making one step towards automated *de novo* ligand design.

In a third part, the afore-mentioned concepts are applied together onto three case studies from the challenging field of protein-protein interactions (PPIs) to illustrate the practicability of the developed approaches in real-life medicinal chemistry examples.

3. Computational details

3.1 Data sets and structure preparation

Protein structures

In all cases, a common workflow for structure preparation was used. The protein structures were centered using openbabel version 2.3.2,[252] and, if present, crystallographic water molecules and buffer molecules were removed. To obtain *apo* structures for the 3D RISM calculations, the ligand was removed. Protein structures were protonated and parametrised using the tleap program from the AMBER18 software package[253] with the ff14SB force field[140] assuming a pH of 7.4. Protonation of ligands was used as provided (PDBbind data set) or carried out using babel assuming a pH of 7.4 (all other structures). For ligand parametrisation, the GAFF force field (version 1.7)[144] was employed. Atomic charges of the ligands were calculated with antechamber using the AM1-BCC charge model[254,255]. All structures used in this work are provided in the Electronic Appendix (with respective exact paths referenced in the respective result sections).

PDBbind refined 2019:

A subset of 3812 structures from the PDBbind refined set 2019[121] was used (list of PDB codes in SI). Due to the large number of complexes, the preparation had to be automatised, and only those structures for which automated preparation was successful were included (since for instance manual parametrisation of ligands, cofactors or special residues was not feasible). Ligand mol2 and sdf files as well as pdb files of the binding pockets were taken as provided.

PDBbind core set 2013:[123]

All complex structures were taken as provided; 18 structures out of the 195 were excluded since they contained covalently bound cofactors or since the systems were too large for feasible calculations (list of used PDB codes in appendix, 7.2). Prior to structure preparation, duplicate chains of multimer complexes were removed.

XIAP, hTEAD, Bcl-xL:

All complex structures were retrieved from the PDB (the codes are given in the respective chapters). For XIAP structures, the cysteine residues coordinating the Zn were renamed to CYM according to AMBER naming conventions. For hTEAD structures, the covalently bound cofactor in the central pocket was removed prior to parametrisation since it is far away from the studied binding site.

Ligand data sets for docking

Starting from SMILES codes, protonation and the generation of a 3D structure as starting point for docking were performed using the RDKit functionalities in KNIME[256] with default settings.

The specific data sets that were used are described in 3.3 and are given in the Electronic Appendix.

3.2 RISM calculations

3D RISM *uv*:

For all 3D RISM calculations, a precomputed solvent susceptibility was used which was calculated in the working group with the dielectrically consistent (DRISM/HNC) theory for pure water (modified SPC/E model[257]) using in-house 1D RISM *vv* code. Respective calculations were performed on a logarithmic grid with 512 grid points with grid spacing ranging between 0.0059 Å and 164.02 Å. The solvent density was set to 0.03334 Å^{-3}, the dielectric constant to 78.4, and the temperature to 298.15 K. The convergence threshold was set to a maximum deviation of 10^{-7} of the direct correlation functions between successive iteration steps.[258] All 3D RISM *uv* calculations were performed with the software developed in the working group on cubic grids with a grid spacing of 0.25 Å and box dimensions based on the protein size (in each dimension, the box length was set to the maximum distance between two protein atoms in this dimension plus 14 Å on each side). The PSE-2 closure was applied throughout. Long range electrostatics were evaluated using the PME of order 8, short range interactions were cut at 14 Å. The convergence threshold was set to 10^{-5}.

3D RISM *uu*:

Solving the *uu* RISM equations of two solutes 1 and 2 in solvent requires prior *uv* calculations of both solute species, in this case a simple, spherical probe and a protein, in the respective solvent. Therefore, consecutive RISM calculations were performed with the software developed in the Kast working group.

The *uv* calculations of the probes (parameters given in Table 1) were carried out as in earlier work[259] with in-house 1D RISM *uv* code (with the convergence criterion set to 10^{-5}) using the modified SPC/E susceptibilities. The *uv* calculations of the proteins were carried out using the in-house 3D RISM code as described above. Afterwards, the *uu* calculations were performed with the in-house 3D RISM *uu* code implemented by F. Mrugalla.[259] As probes, an uncharged c3 probe as well as charged n4 and o probes were used with epsilon and sigma LJ parameters derived from the respective values of the atom types c3, n4, and o in the GAFF force field. The charges of the n4 and o probe were set to +1 and –1, respectively. In this work, 3D RISM *uu* calculations were performed for the complexes in the PDBbind core set and for selected structures of the proteins XIAP, Bcl-xL, and hTEAD (pdb codes given in the respective chapters).

Table 1: Used charge, epsilon and sigma parameters for the uu probes.

probe	charge / e	sigma / Å	epsilon / 10^{-21} J
c3	0	3.39967	0.760078
n4	+1	3.25000	1.181109
o	-1	2.95992	1.459017

3.3 Docking experiments

Docking experiments were carried out using GOLD (version 18.1 for pose recovery and 20.1 in all other cases).[51] In all cases, the binding site was determined from the bound ligand in the complex structure with a radius of 10.0 Å. The options "flip_free_corners", "match_ring_templates" and "flip_planar_n" were enabled; for virtual screening runs, also the option "allow early termination" was enabled. After docking, the obtained poses were converted to pdb files using babel for further post-processing.

Pose recovery:

For pose recovery experiments, 100 diverse poses (RMSD > 1.5 Å) were generated for each ligand of the PDBbind core set 2013 three times, resulting in 300 poses per ligand.

Virtual screening of XIAP DUD-E data set:

For virtual screening on the protein XIAP, the respective DUD-E data set was used.[127] The molecules were taken as provided and docked into the provided pdb structure 3hl5.[262] A maximum of 25 poses was generated, and the top solution was kept.

The ROC-AUC values for the resulting rankings were determined using the ROC-node in KNIME.[256]

Virtual screening of fragments in XIAP:

For virtual screening of fragments, the library as described by Sandór *et al*. was used,[260] with three additional fragments derived from the ligand bound in the complex structure 5c7a.[261] Starting from the SMILES, the molecules were prepared as described in 3.1 and docked into pdb structure 3hl5.[262] A maximum of 25 poses was generated, and the top solution was kept.

Docking of hTEAD4 compounds provided by the Brunschweiger group:

The hTEAD4 data set provided by the Brunschweiger group was prepared as described in 3.1, starting from the SMILES, and docked into structure 6q36. The Cl-indole moiety of the co-crystallised modified YAP-peptide was used as scaffold constraint and for the binding site definition. The search efficiency was set to 200 % owing to the size and number of rotatable bonds of the molecules.

3.4 Analysis of water thermodynamics

Water placement:

Placement of distinct water molecules was performed as described in earlier work[192] based on the pair distribution function of the water oxygen atom, $g_{\mathrm{O}}(\mathbf{r})$. In the placement algorithm, only volume elements with a $g_{\mathrm{O}}(\mathbf{r})$ value exceeding the 99.9 percentile threshold are retained. These elements are kept in a list and sorted descending by their $g_{\mathrm{O}}(\mathbf{r})$ values. Water positions are then calculated by iterating over this list: The volume element with the highest $g_{\mathrm{O}}(\mathbf{r})$ value is saved as a water position $\mathbf{r}_w$, then this element and all other elements within a distance cutoff of 2.5 Å are deleted from the list as they are considered to belong to the same water molecule. This procedure is repeated until the list is empty.

Analysis of free energy density:

The hydration free energy density $\rho_{\mathrm{G}}(\mathbf{r})$ represents the contribution of spatial regions around the protein to its total free energy of hydration, $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$. However, this field is rather rugged, so that small variations of a given position $\mathbf{r}_i$ can result in a large difference of the resulting hydration free energy value. To overcome this, a Gaussian convolution of the $\rho_{\mathrm{G}}(\mathbf{r})$ field was performed using a sigma of 1.4 Å (representing approximately a water molecule's radius), which formally corresponds to an integration with smooth boundaries and results in a respective smoothed field $\rho'_{\mathrm{G}}(\mathbf{r})$. To determine the individual contribution of a specific water molecule w to the total free energy of hydration of the protein, the $\rho'_{\mathrm{G}}(\mathbf{r})$ field was evaluated at the respective position $\mathbf{r}_w$. For conversion of this respective free energy density value $\rho'_{\mathrm{G}}(\mathbf{r}_w)$ to an absolute energy contribution of the respective water molecule, $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w}$, $\rho'_{\mathrm{G}}(\mathbf{r}_w)$ was multiplied with an arbitrary volume of 11.494 Å^3, corresponding to a spherical water molecule with a radius of 1.4 Å (which was used as sigma in the respective Gaussian convolution). Since only selected water positions were placed within this workflow (see above, 99.9 percentile threshold w.r.t. g_{O}-function), it has to be noted that the sum of all individual $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w}$ values of all placed water molecules w is in this case not equal to the total value of μ_{ex} that is obtained by integration of the $\rho_{\mathrm{G}}(\mathbf{r})$ field, i.e.:

$$\mu^{\mathrm{ex}}_{\mathrm{PSE}\text{-}n} = \Delta_{\mathrm{hyd}}G_{\mathrm{P}} \equiv \int \mathrm{d}\mathbf{r}\rho_G(\mathbf{r}) = \int \mathrm{d}\mathbf{r}\rho'_G(\mathbf{r}) \neq \sum_w \Delta_{\mathrm{hyd}}G_{\mathrm{P},w} \tag{68}$$

Similar to obtaining the individual contribution $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w}$ at a specific water position, the respective field can also be evaluated at the positions where ligand atoms are present in the *holo* complex structure. Thus, each ligand atom l is assigned a value $\Delta_{\mathrm{hyd}}G_{\mathrm{P},l}$ which is a measure for the $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contribution of the *apo* water molecule that this ligand atom replaces upon binding. Negative values accordingly denote displacement of a stable, favourably bound water molecule, positive values that of a water molecule with an unfavourable contribution to $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$. This procedure allows for a direct and fast mapping of *apo* water thermodynamics onto *holo* ligand atom positions and thus enables an efficient correlation between ligand chemistry and water thermodynamics.

Local empirical corrections for 3D RISM:

In a proof-of-concept study, the local V_{m}- and q-based correction as defined in Eq. (66) was carried out for an exemplary complex within the PDBbind refined set (pdb: 2xbv) with $c_V = -0.10251$ kcal mol^{-1}Å^{-3} and $c_q = 15.728$ kcal mol^{-1} e^{-1} as determined by Tielker *et al.*[198]. It has to be noted that these values were obtained for an EC-RISM based workflow (at the MP2/6-311+G(d,p)/EC-RISM// B3LYP/6-311+G(d,p)/PCM level of theory); however, they can serve as a first reference in the provided proof-of-concept study. For each water position $\mathbf{r}_w$ as determined from the $g_{\mathrm{O}}(\mathbf{r})$ field, the respective individual contribution to the total free energy of hydration of the protein, $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w}$, was determined by summation of the $\mu^{\mathrm{ex}}(\mathbf{r}_i)$ values of all volume elements within 2.5 Å of the water position $\mathbf{r}_w$, thus yielding two values, $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w,\mathrm{orig}}$ and $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w,\mathrm{corr}}$, for each water molecule.

3.5 Analysis of probe densities

With help of 3D RISM *uu* calculations, the pair distribution function of selected pharmacophoric probes can be determined in a protein binding site. Based on the respective 3D RISM *uv* calculation of a given protein in water, three individual 3D RISM *uu* calculations were performed for each of the three pharmacophoric probes (c3, n4, o, s. Table 1), resulting in three individual densities fields. These fields can be evaluated at the position of ligand atoms l in docking poses or experimental complex structures (linear interpolation from the nearest grid points). The obtained value $g_{l,p}$ hence corresponds to the g-function value of the respective probe (c3, n4, or o) in the *apo* binding site at the position where the ligand atom l is located in the complex. These values allow to assess how well the respective ligand structure matches the binding site properties, following the assumption that for instance ligand oxygen atoms should be located in areas of the binding site with high g-function values for the o probe. To achieve a quantitative measure for such a ligand structure – probe density match, a respective score was developed with the aim to capture how well the ligand atoms are in line the probe fields. The respective equations are discussed in the results section in 4.2.2.

3.6 MD Simulation and analyses on Bcl-xL

In the following, the workflow for the MD simulation and clustering of Bcl-xL structures is described which was performed by J. Borchert within the scope of a bachelor thesis that led to the structures which were used in this work. The work described here was not done by the author but is described here for clarification and is strongly oriented at the methods section in the respective thesis.[310]

All systems simulated in J. Borchert's work were parameterised using the tleap program with the ff14SB force field[140] for proteins and GAFF[144] for ligands. The partial charges of ligands were calculated using antechamber and the AM1-BCC charge model.[254,255] Single-charged ions were added to neutralise the system (Na^+ and Cl^-); they were parameterised using the parameters of Li and Merz.[263] The box dimensions were chosen to have at least 18 Å distance between the protein and the edge of the box in each dimension. SPC/E water was used as the water model.

The starting point of the simulations was the NMR structure of Bcl-xL in complex with a co-crystallised ligand from Abbot (pdb: 1ysi[307]). For the simulation, the acylsulfonamide in the Abbot ligand was assumed to be deprotonated since acylsulfonamides are known in literature as isosteres for carboxylates with pK_a values in the range of 4.[264] First, the Bcl-xL complex with the Abbott ligand was simulated, and the obtained trajectory was clustered using the Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm with the program cpptraj.[265,266] The parameters used for clustering are shown in Table 2.

Table 2: Parameters of the DBSCAN clustering of the investigated systems. The parameter minPt is the minimum number of structures per cluster, and the parameter σ defines the distance (RMSD values) in Å between the structures in a cluster.

System	*minPt*/*σ*
Bcl-xL + Abbott ligand	100/2
Bcl-xL + Ugi model compound 1	50/2,2
Bcl-xL + Ugi model compound 2	50/2,2

The centroid of the highest populated cluster was then split into protein and ligand. Then, the two Ugi model compounds provided by the Brunschweiger group (structures given in 4.3.2) were docked into the binding site of this representative structure using the program GOLD (version 18.1).[51] Here, the binding site of Bcl-xL was defined based on the location of the Abbott ligand with a radius of 10.0 Å, and the biphenyl moiety served as a scaffold constraint for the biphenyl moieties of the Ugi model compounds. For each of the two ligands, 100 poses were created using the ChemPLP scoring function. From these poses, the top ranked structure was visually inspected and selected as the initial structure for the simulations. These newly generated complexes were prepared in the same way as described above and then simulated, followed again by clustering to obtain a representative structure for the trajectories of the complexes with the two Ugi model compounds. This workflow is illustrated in Figure 52 in the respective results chapter 4.3.2.

All simulations were performed at 298.15 K at a pressure of 1 bar, and the size of a single time step was always 2 fs. The simulations were performed using Amber 18.[253] First, energy minimisation was performed, then the system was heated to a temperature of 298.15 K in 20 ps in the NVT ensemble using a Langevin thermostat.[151] For this part of the simulation, a harmonic restraint with a spring constant of 4.00 $k_\mathrm{B}T$ was placed on the $C_\propto$ atoms. In the third part, the system was brought to the desired pressure in the *NpT* ensemble. This was done over a period of 4 ns using a Berendsen barostat[152] using the same restraint as in the previous step was used. The system was then simulated in the *NpT* ensemble. This procedure was performed for all simulated systems. The parameters of each simulated system can be found in Table 3.

Table 3: Simulated systems with charges of protein and ligand, q_P and q_L, simulation time t, and number of atoms.

system	q_P /e	q_L /e	t / ns	number of atoms
Bcl-xL + Abbott ligand	-12	-1	300	73564
Bcl-xL + Ugi model compound 1	-12	0	300	62282
Bcl-xL + Ugi model compound 2	-12	0	330	62279

Using cpptraj, the ions and water molecules were first removed from the resulting trajectories, and the complex was centered on its origin. The modified trajectories were saved and served as the basis for all further analyses.

For this work, the original pdb structure 1ysi, the docking poses of both Ugi model compounds in Bcl-xL, and the three respective representative structures obtained by the clustering for the systems Bcl-xL + Abbot ligand, Bcl-xL + Ugi model compound 1, and Bcl-xL + Ugi model compound 2 were used as input structures for RISM calculations.

3.7 Visualisation

All graphical representations of protein and ligand structures in this work were generated using PyMOL version 1.8.[267] Histograms and scatter plots were generated using R version 3.4.4.[268] In the histograms, values on the y-axis correspond to probability densities (with reciprocal units of the parameter on the x-axis), such that the total area of the histograms always has a value of 1.

4. Results and Discussion

4.1 Analysis of water thermodynamics

The first part of this work focuses on the thermodynamic properties of binding site water molecules and their relevance for drug design. To obtain robust and meaningful results, a large-scale analysis was performed on several thousand complex structures of the PDBbind refined set. Using the approaches described in 3.4, the positions and $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions of hydration sites in the respective binding sites were calculated both for the respective *apo* (i.e. without the bound ligand) and *holo* form.

In chapter 4.1.1, an in-depth analysis w.r.t. the *apo* water thermodynamics is presented. At first, parameters like experimental B-factors and the protein microenvironment are correlated with water "happiness" to elucidate general characteristics of "happy" and "unhappy" water molecules. Afterwards, the influence of water thermodynamics on aspects like replaceability and druggability is investigated. Finally, a detailed analysis follows that correlates the replacement of "happy" and "unhappy" water molecules by specific functional groups with respective binding free energies, thus allowing to derive rules about which kind of binding site water molecules should be preferentially replaced with which kind of ligand groups.

In chapter 4.1.2, the thermodynamic properties of *holo* hydration sites are analysed in a similar manner, and an attempt is made to correlate the presence of "unhappy" water molecules in complexes with ligand affinity.

In chapter 4.1.3, the findings from the analyses in 4.1 and 4.2 are exemplified on suitable sets of MMPs to provide concrete illustrations of how the 3D RISM-based water placement and characterisation can help to optimise a given ligand.

Besides, a brief poof-of-concept study is presented in chapter 4.1.4 in which the well-established PMV-correction for 3D RISM (s. 2.3.3) is applied locally on grid-level. While this correction does not change the overall picture significantly (as will be shown), such developments could be pursued in the future to further advance 3D RISM-based hydration site characterisation.

All raw data for the analyses in 4.1 can be found in the Electronic Appendix (Electronic_Appendix/ PDBbind_refined_set/). This includes respective ligand and protein structures (ligand.pdb, pocket.pdb) as well as the calculated water positions with thermodynamic properties (Ghyd@water_apo.pdb, Ghyd@water_holo.pdb) and the interpolated *apo* water thermodynamic data on the ligands (Ghyd@lig.pdb) for each structure within the used PDBbind refined subset.

4.1.1 Analysis of *apo* water thermodynamics

4.1.1.1 Reproduction of experimental water positions

For the more than 3800 structures of the used PDBbind refined subset, the positions (and individual $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions) of *apo* water molecules were predicted based on 3D RISM calculations and the methodology presented in 3.4. It should be noted that, in the context of this work, the term “water position” always refers to the predicted position of the water oxygen atom since only the respective g_{O}-function is considered in the placement.

The correct placement of water molecules is the prerequisite for any further analysis; therefore, it was first evaluated if the water positions predicted by the used algorithms come close to the respective experimental water positions. Indeed, the distribution of the distances between the available experimentally determined water positions and the corresponding nearest predicted water positions (Figure 2) shows good agreement. For 88 % of the predicted *apo* water molecules in the binding sites of the used PDBbind refined subset, the distance to the nearest predicted water position is below 2.0 Å; for 74 % and 49 % below 1.5 Å and 1.0 Å, respectively.

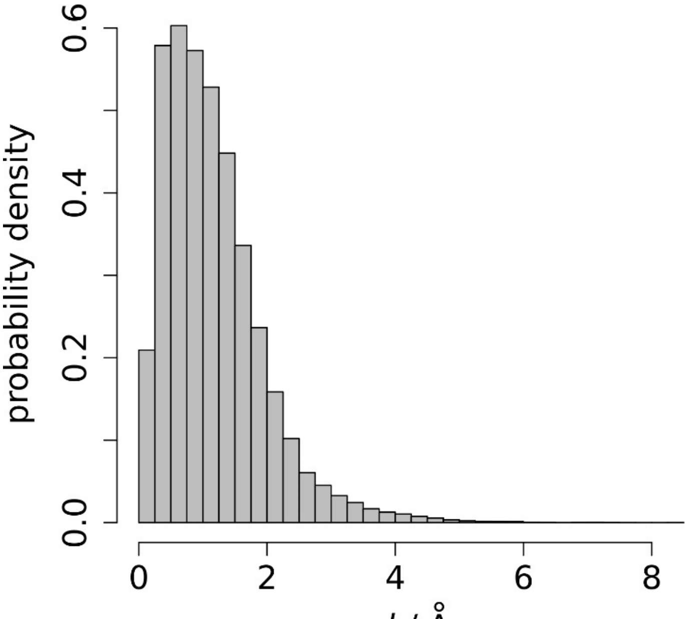

Figure 2: Probability densities of the distances d between the experimental holo water positions and the corresponding nearest calculated apo water positions as predicted by 3D RISM-based algorithms for all structures within the used PDBbind refined subset. The probability density has the inverse unit of the x-axis parameter, i.e. 1/Å. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/Figure2_distances_apo/).

As will be shown later in this work, some discrepancies can be attributed to the fact that calculated *apo* water positions are compared with experimental water positions of *holo* complex structures; when considering the ligand, even better overall agreement can be achieved (chapter 4.2).

Besides, it has to be noted that the protein structures used in the analysis are no “true” *apo* structures but “pseudo” *apo* structures that were generated by removing the ligand from the complex structures, which of course is an approximation. Yet, previous work[192] already showed that this procedure leads to reliable results, with predicted hydration sites that overlap nicely with experimental water positions in several different fXa X-ray structures and that allow to explain SAR trends. Nevertheless, the potential influence of this approximation was investigated on selected examples of the PDBbind refined set for which “true” *apo* structures are readily available in the Protein Data Bank:[269] Figure 3 shows the superposition of the calculated water positions (based on the respective “pseudo” *apo* structures) with the X-ray water positions of “true” *apo* structures for HIV-1 protease, neuraminidase, carbonic anhydrase, and fXa. For all examples, the calculated “pseudo” (cyan) and experimental “true” *apo* water positions (red) are close, with only small deviations that can be attributed to the fact that two different X-ray structures of the same protein always show certain deviations, e.g. different loop or sidechain conformations or the presence of buffer molecules, even if two *apo* structures are compared. This again suggests that the use of “pseudo” *apo* structures results in reliable hydration site predictions comparable to those on “true” *apo* structures.

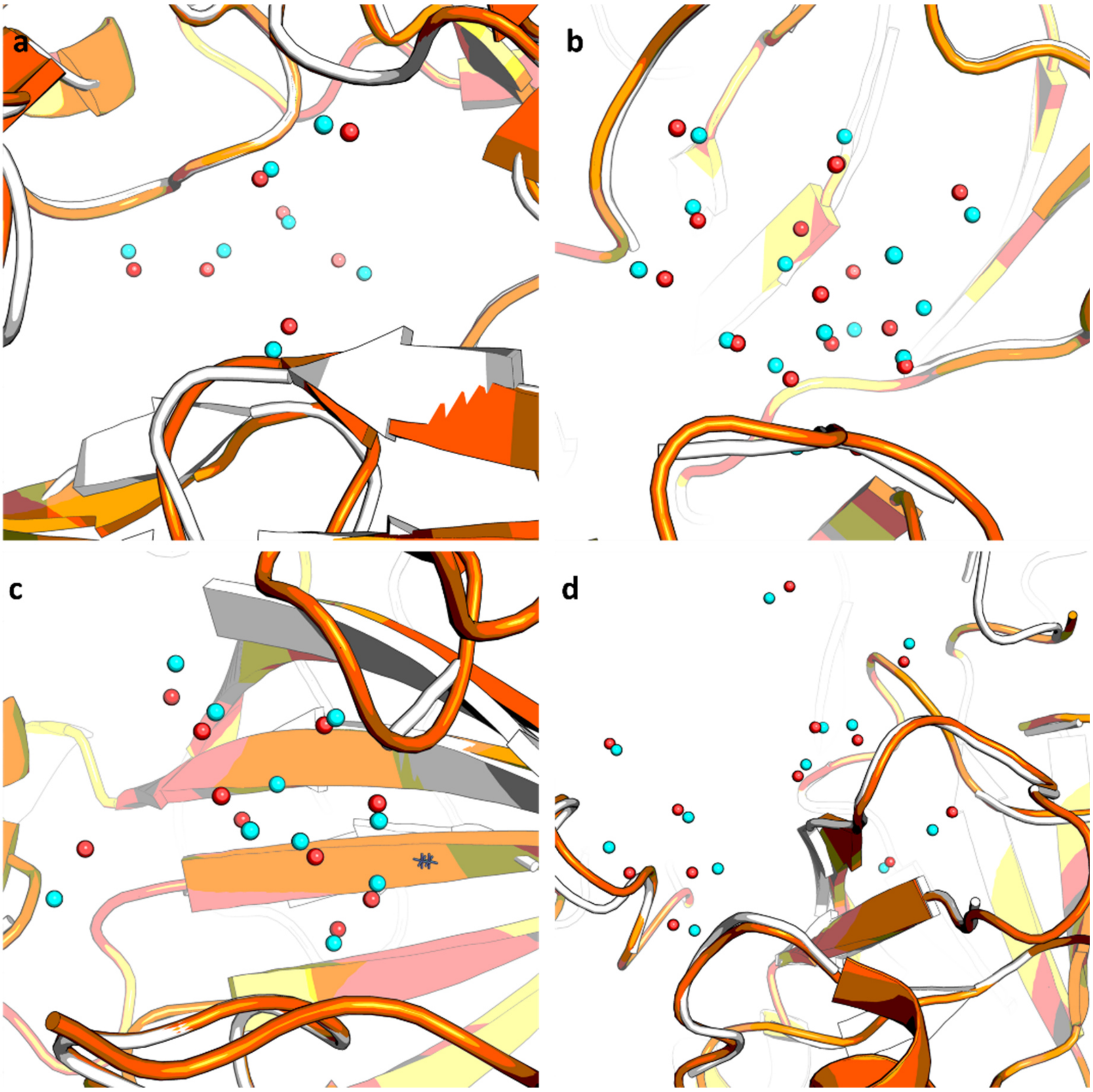

Figure 3: Comparison of apo water positions as calculated by the 3D RISM-based algorithms (cyan) using "pseudo" apo structures (white) with experimental water positions (red) of "true" apo structures (orange) of the same protein for exemplary structures in the used PDBbind refined subset. All experimental water positions of the "true" apo structure (aligned to the used "pseudo" apo structure using Pymol) within 4 Å of the holo ligand are shown together with the respective nearest corresponding calculated apo water position as predicted by the algorithms described in 3.4. a) HIV-1 protease (1hbv,[270] 3ixo[271] ("true" apo)); b) neuraminidase (1f8b,[272] 6d3b[273] ("true" apo)), c) carbonic anhydrase (3ibu,[274] 1ca2[275] ("true" apo)); d) fXa (2xbv[276], 1hcg[277] ("true" apo)). The respective structures can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/structures/) in the respective structure folders (1hbv, 1f8b, 3ibu, 2xbv).

4.1.1.2 Are water molecules with low B-factors better reproduced?

For the experimentally determined water molecules, structural B-factors are available, which are a measure for the molecules' thermal mobility. Studies suggest that conserved water molecules tend to have lower B-factors.[213] Therefore, it was investigated if water molecules with lower B-factors are better reproduced by the 3D RISM-based calculations (Figure 4, Table 4; water molecules within 1.5 Å of the ligand were excluded since the experimental B-factor might be influenced by the presence of the ligand which was not considered here).

Indeed, the positions of the 10 % and 25 % most localised water molecules are reproduced within 1.0 Å in 73 % and 69 % of cases, within 2.0 Å in 97 % and 96 %, respectively. On the other hand, only 28 % and 32 % of the 10 % and 25 % of water molecules with the highest B-factors can be reproduced within 1.0 Å (79 % and 76 % for a 2.0 Å threshold). A similar trend was already observed in an analysis of predicted water molecules at protein-protein interfaces.[222] Apart from high thermal mobility, high B-factors can be the result of disorders in the crystal structure or a not well-defined electron density. For some cases, this might explain the discrepancies between predicted and experimental water positions.

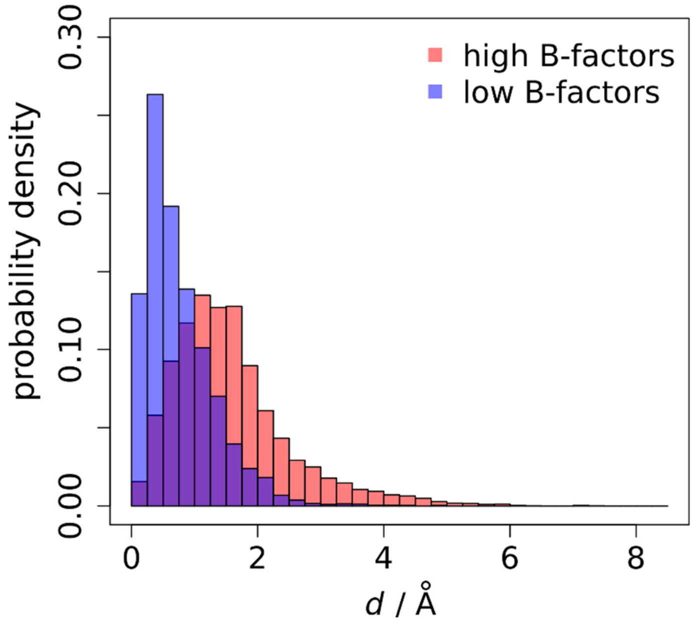

Figure 4: Probability densities of the distances between the experimental holo water positions and the corresponding nearest calculated apo water positions as predicted by 3D RISM-based algorithms for subsets of experimental water positions within the used PDBbind refined subset w.r.t. B-factor (blue: 10 % lowest B-factors, red: 10 % highest B-factors). The probability density has the inverse unit of the x-axis parameter, i.e. 1/Å. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure4_Table4_Bfactors_apo/).

Table 4: Percentages of the experimental water positions (holo, since PDBbind refined set only contains complexes) in complexes of the used PDBbind refined subset that are correctly reproduced by the calculated apo water positions by the 3D RISM-based placement algorithm (using three different distance thresholds, 1.0, 1.5 and 2.0 Å). Percentages are shown for all experimental water molecules ("all") as well as for subsets of water molecules with the highest and lowest B-factors ("min X %" and "max % B-factor"). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure4_Table4_Bfactors_apo/).

dist. threshold / Å	all	min 10 % B-fact.	min 25 % B-fact.	max 25 % B-fact.	max 10 % B-fact.
1.0	49.1 %	73.0 %	68.5 %	31.8 %	28.3 %
1.5	73.5 %	90.1 %	87.6 %	59.0 %	54.5 %
2.0	87.8 %	96.5 %	95.7 %	79.0 %	76.2 %

4.1.1.3 Distribution of water thermodynamics within the data set

In the following sections of this chapter, different analyses are presented w.r.t the thermodynamic properties of the predicted *apo* water molecules in the structures of the used PDBbind refined subset - in this work, this always refers to the predicted *apo* water molecules' individual $\Delta_{\text{hyd}}G_{\text{P}}$ contributions as calculated by the 3D RISM-based algorithm as described in 3.4. As in indicated in the introduction in 2.3.3.3, empirical correction terms were developed within the author's working group that improve the calculation of absolute solvation free energy values. A localised version of this empirical correction was also applied on a selected case study within this work (in detailed presented in 4.1.4), but the analysis revealed that such a correction has no significant effect on the individual $\Delta_{\text{hyd}}G_{\text{P}}$ contributions of specific water molecules. Hence, all $\Delta_{\text{hyd}}G_{\text{P}}$ contributions discussed here do not include any further empirical corrections.

Before correlating $\Delta_{\text{hyd}}G_{\text{P}}$ contributions with specific parameters – like the properties of near residues or replacing ligand groups -, it is reasonable to get an overview about the general distribution of the water molecules' $\Delta_{\text{hyd}}G_{\text{P}}$ contributions within the data set. In Table 5, respective average $\Delta_{\text{hyd}}G_{\text{P},w}$ values over the whole dataset are given together with median and percentile values; the respective histogram is shown in Figure 5. As could be expected, the majority of water molecules has $\Delta_{\text{hyd}}G_{\text{P}}$ contributions which are slightly favourable or unfavourable (with a median of -0.08 kcal·mol^{-1}) and fewer water molecules with high absolute $\Delta_{\text{hyd}}G_{\text{P},w}$ values. However, it can be seen that the distribution is not perfectly symmetric but that there are slightly more water molecules with highly favourable than with highly unfavourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions, and they also have slightly higher absolute values. This small, inherent bias towards more negative, and hence favourable, $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions has to be kept in mind in the upcoming analyses.

Table 5: Average $\Delta_{hyd}G_{P}$ contributions (in kcal/mol) as well as respective median and percentile values for the predicted apo water molecules in all structures in the used PDBbind refined subset as calculated by the 3D RISM-based algorithms described in 3.4. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure5_Table5_Ghyd_distribution/).

$\Delta_{\mathrm{hyd}}G_{\mathrm{P},w}$ / kcal·mol^{-1}					
average	median	10 % percentile	25 % percentile	75 % percentile	90 % percentile
-0.48 ± 2.30	-0.08	-2.15	-0.78	0.31	0.76

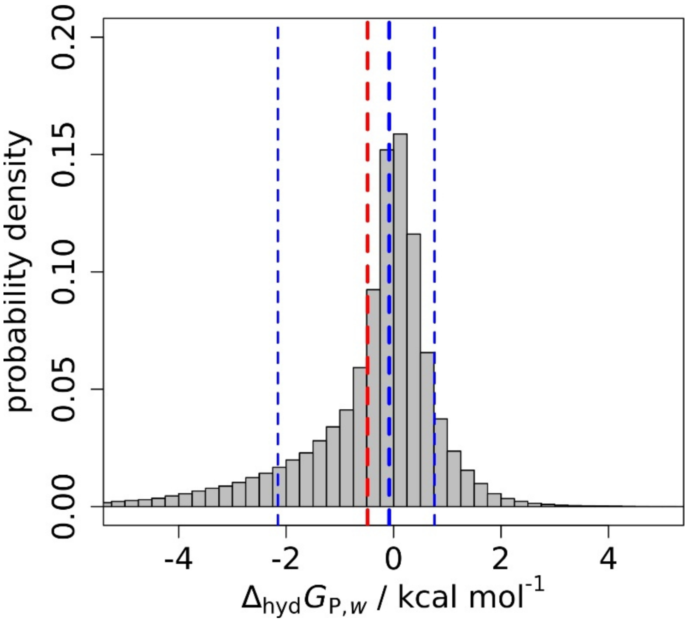

Figure 5: Probability densities of the calculated $\Delta_{hyd}G_{P}$ contributions (in kcal/mol) of all predicted apo water molecules in the used PDBbind refined subset as calculated by the algorithms presented in 3.4. The average value is shown as a red dashed line; the median, 10 %, and 90 % percentile are shown as blue dashed lines. The probability density has the inverse unit of the x-axis parameter, i.e. kcal^{-1}·mol. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure5_Table5_Ghyd_distribution/).

4.1.1.4 Are localised water molecules "happier"?

Based on the B-factor analysis in 4.1.1.2, the next aim was to investigate the reasonable hypothesis that water molecules with low B-factors may be highly localised because they are tightly bound and undergo favourable interactions with their protein environment. This would imply that water molecules with lower B-factors have a more favourable contribution to $\Delta_{\text{hyd}}G_{\text{P}}$. On the other hand, they could also be trapped in an unfavourable environment and could have unfavourable entropic properties.

To investigate this, it was analysed if there is a correlation between the B-factors of the experimental water positions and the $\Delta_{\text{hyd}}G_{\text{P}}$ contributions of the corresponding predicted *apo* water molecule positions for the structures in the used PDBbind refined subset. In this analysis, badly predicted water molecules with a distance between experimental and predicted position exceeding 1.5 Å were excluded. Besides, experimental water positions in direct proximity of the ligand (3.0 Å) were excluded in this analysis since the water positions and $\Delta_{\text{hyd}}G_{\text{P}}$ contributions were calculated for the *apo* binding site, and the experimental *holo* B-factors might be highly influenced by the presence of the ligand.

In Figure 6, the distribution of B-factors is shown for two subsets of experimental water positions based on the $\Delta_{\text{hyd}}G_{\text{P}}$ contributions of their respective nearest calculated water positions (25 % most and least "happy" water molecules). In addition, in Table 6, the respective average B-factors of different subsets are given.

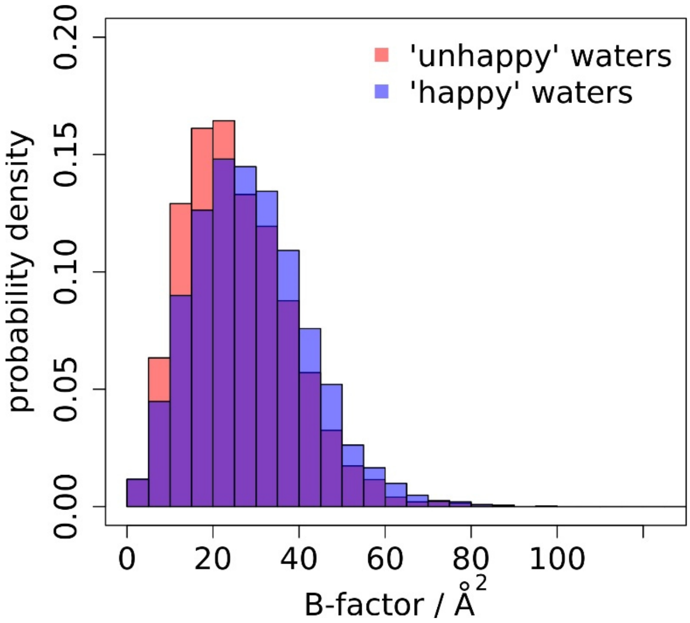

Figure 6: Probability densities of the B-factors of the experimental water positions within the used PDBbind refined subsets for subsets of experimental water molecules w.r.t. the $\Delta_{hyd}G_{P}$ contributions of the corresponding nearest predicted water molecules as predicted by 3D RISM (blue: water molecules with the 10 % most favourable $\Delta_{hyd}G_{P}$ contributions, red: water molecules with the 10 % least favourable $\Delta_{hyd}G_{P}$ contributions). The probability density has the inverse unit of the x-axis parameter, i.e. 1/Å^2. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure6_Table6_Bfactor_Ghyd/).

Table 6: Mean B-factor values of experimental water positions for water subsets based on the $\Delta_{hyd}G_P$ contributions of the corresponding predicted water positions as determined from 3D RISM for complexes in the PDBbind refined set (four subsets containing only those water molecules with the 10 %, and 25 % most and least favourable $\Delta_{hyd}G_P$ contributions). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ data/ PDBbind_refined_set/ Figure6_Table6_Bfactor_Ghyd/).

	X % most favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions		X % least favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions	
	10 %	25 %	25 %	10 %
B-factor	28.8 ± 13.3	29.2 ± 13.0	26.9 ± 13.2	25.9 ± 13.3

Intriguingly, there are only small differences between the subsets, and in contrast to expectations, even a trend for slightly lower B-factors for “unhappy” waters can be observed, suggesting that many highly localised water molecules are rather “unhappy”. Thus, this analysis shows that the experimental B-factor alone does not provide information about a hydration site’s thermodynamic properties and that advanced theoretical methods like 3D RISM are needed for such a characterisation.

4.1.1.5. Influence of the protein microenvironment

A water molecules’ thermodynamic properties are determined by its microenvironment, i.e. near binding site residues and other solvent molecules. Other studies[212] have shown that the atomic density, defined as the number of protein heavy atoms within 3.5 Å distance of a water molecule, is an important property to distinguish between bound and replaced water molecules. In this work, the atomic density as well as the apolar and polar atomic density (here simply defined as the number of carbon or oxygen, nitrogen, and sulphur atoms within the defined distance) was analysed for predicted water molecules in complexes of the used PDBbind refined set, divided into subsets based on their $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions.

The average total, polar, and apolar atomic densities for water subsets containing the water molecules with the 10 % and 25 % most and least favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions are given in Table 7; in addition, the respective histograms for the 10 % subsets are shown in Figure 7. The results show that the total atomic density is slightly higher for water molecules with less favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions, suggesting that they are more buried. A more pronounced trend can be observed for the apolar atomic density, which is in average almost twice as high for “unhappy” water molecules than for “happy” ones. The polar atomic density, on the other hand, shows the opposite trend, albeit less pronounced. These findings are highly intuitive since water molecules can undergo hydrogen bonds with polar environment, which is not possible in deeper and more hydrophobic pockets in the protein.

Table 7: Average total, polar, and apolar atomic density values (i.e the number of i) protein heavy atoms, ii) protein nitrogen, oxygen, and sulphur atoms, or iii) protein carbon atoms within 3.5 Å around the water position as predicted by 3D RISM-based algorithms) for water molecule subsets based on $\Delta_{hyd}G_P$ contributions (i.e. comprising only those water molecules with the X % most and least favourable Δ_{hyd}G$_P$ contributions) for the complexes of the used PDBbind refined subset. All water molecules, not only binding site water molecules, were considered for the analysis. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure7_Table7_atomic_density/).

	X % most favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions		X % least favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions	
atomic density	10 %	25 %	25 %	10 %
total	3.5 ± 1.6	3.4 ± 1.6	4.1 ± 2.2	4.7 ± 2.5
polar	2.1 ± 1.0	1.9 ± 1.0	1.5 ± 1.2	1.7 ± 1.3
apolar	1.4 ± 1.2	1.4 ± 1.2	2.6 ± 1.7	3.0 ± 1.9

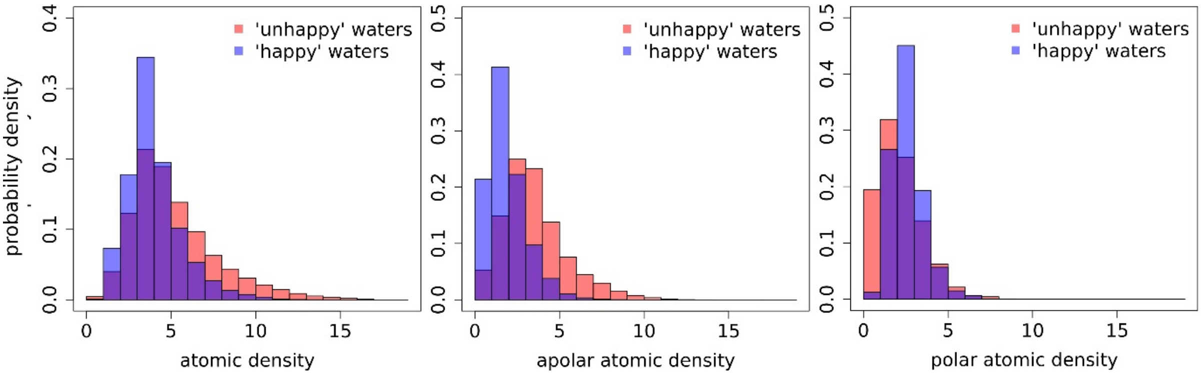

Figure 7: Probability densities of the total atomic density (left), apolar atomic density (middle), and polar atomic density (right) for subsets of water molecules with the 10 % most ("happy", blue) and least ("unhappy", red) favourable $\Delta_{hyd}G_P$ contributions in the used PDBbind refined subset. The probability density has the inverse unit of the x-axis parameter, i.e. 1/atomic density, 1/apolar atomic density, and 1/polar atomic density, respectively. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure7_Table7_atomic_density/).

In addition, also the number of other water molecules and of charged groups (Glu/Asp and Lys/Arg sidechain groups) in the proximity (3.5 Å) of water molecules in the two subsets (Figure 8, Table 8) was investigated. In accordance with the finding that "unhappy" water molecules tend to be more buried, they have fewer neighbouring water molecules than those of the "happy" subset. W.r.t. the number of charged groups, it can be seen that practically no “unhappy” water molecules are located in the proximity of such groups.

Table 8: Average number of water molecules and charged groups (Asp, Glu, Lys, and Arg sidechains) within 3.5 Å around the water position as predicted by 3D RISM) for water molecule subsets based on $\Delta_{hyd}G_P$ contributions (i.e. comprising only those water molecules with the X % most and least favourable $\Delta_{hyd}G_P$ contributions) for the complexes of the used PDBbind refined subset. All water molecules, not only binding site water molecules (within 3.5 Å of the ligand), were considered for the analysis. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure8_Table8_near_contacts_Ghyd/).

	X % most favourable $\Delta_{\text{hyd}}G_P$ contributions		X % least favourable $\Delta_{\text{hyd}}G_P$ contributions	
	10 %	25 %	25 %	10 %
waters	2.6 ± 1.3	2.5 ± 1.2	2.1 ± 1.2	2.0 ± 1.2
basic/acidic	0.9 ± 0.9	0.7 ± 0.8	0.1 ± 0.4	0.1 ± 0.4

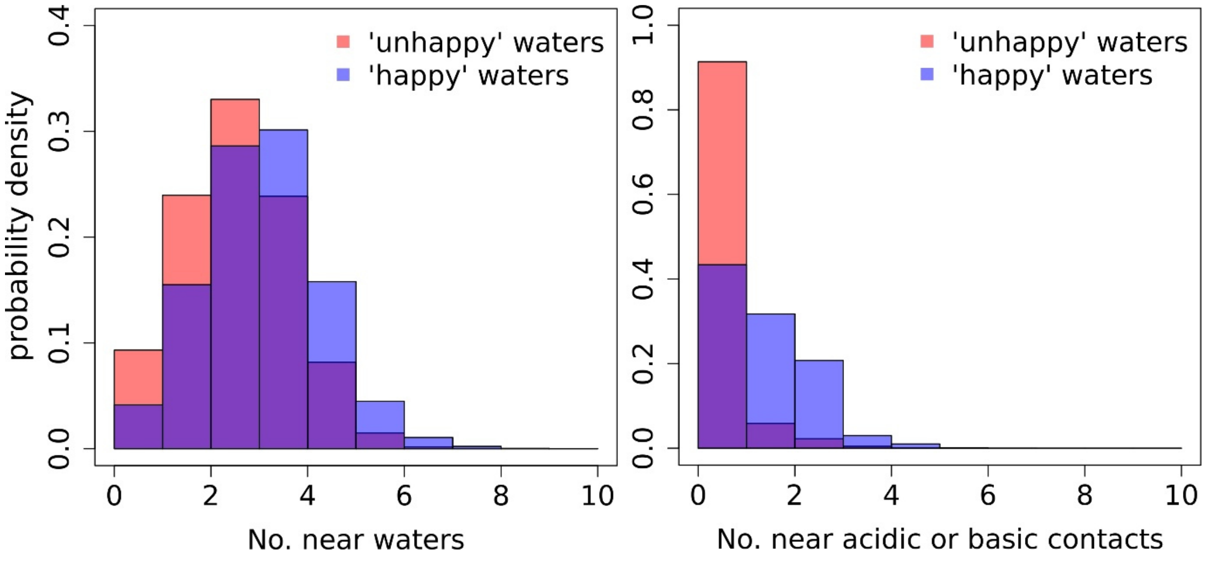

Figure 8: Probability densities of the number of other water molecules and charged groups in the proximity (3.5 Å) of predicted water positions for subsets of water molecules with the 10 % most (“happy”, blue) and least (“unhappy”, red) favourable $\Delta_{hyd}G_P$ contributions in the used PDBbind refined subset. The probability density has the inverse unit of the x-axis parameter. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure8_Table8_near_contacts_Ghyd/).

Following these results, the specific influence of certain protein functional groups was analysed, namely the influence of mainchain amides, the carboxylates of Glu and Asp, the primary amine of Lys, the guanidinium group of Arg, the sidechain amides of Gln and Asn, the hydroxyl groups of Ser, Thr, and Tyr, the aromatic atoms of Phe, Tyr, and Trp, and the aliphatic sidechain atoms of Ala, Leu, Ile, and Val. The average $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions of water molecules in the proximity (3.5 Å) of the specified groups are given in Table 9; the respective histograms are shown in Figure 9.

The results show that water molecules in the proximity of charged groups, i.e. Glu, Asp, Lys, and Arg sidechains, have the most favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions, followed by polar hydroxyl groups and sidechain and main chain amides. Only water molecules near aromatic and aliphatic sidechains show an unfavourable average $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contribution. The observed trends are intuitive and in agreement with a similar study by Beuming *et al.*[278] who analysed the structures of 27 proteins using WaterMap. While the absolute free energy values are shifted to more negative values in this work, the authors observed very similar trends, with water molecules near carboxylates, Lys, and Arg sidechains having the most favourable contributions, followed by hydroxyls, sidechain amides, and backbone carbonyl. In the study by Beuming *et al.*, the least favourable contributions are observed for aromatic and aliphatic groups and, interestingly, for backbone amides. The latter trend is not observed in this work, where water molecules near mainchain nitrogen and oxygen atoms in average show slightly favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions.

Table 9: Average $\Delta_{hyd}G_{P}$ contributions (in units of kcal/mol) of water molecules in the proximity (3.5 Å) of specific protein groups (carboxylates of Asp and Glu, amine of Lys, guanidinium of Arg, hydroxyls of Ser and Thr, sidechain amides, main chain O atoms, main chain N atoms, aromatic atoms in Trp, Phe, and Tyr, aliphatic atoms in Ala, Leu, Ile, Val) for complexes in the used PDBbind refined subset. All predicted water molecules were considered in this analysis. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure9_Table9_Ghyd_AA/).

Functional Group	av. $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w}$ / kcal·mol^{-1}
Carboxylates	-2.02 ± 4.80
Lys Amine	-1.89 ± 2.87
Arg Guanidinium	-1.36 ± 2.47
Hydroxyls	-0.64 ± 3.66
Sidechain Amides	-0.38 ± 1.58
Main chain O	-0.34 ± 1.41
Main chain N	-0.15 ± 1.57
Aromatic	0.08 ± 1.37
Aliphatic	0.19 ± 1.15

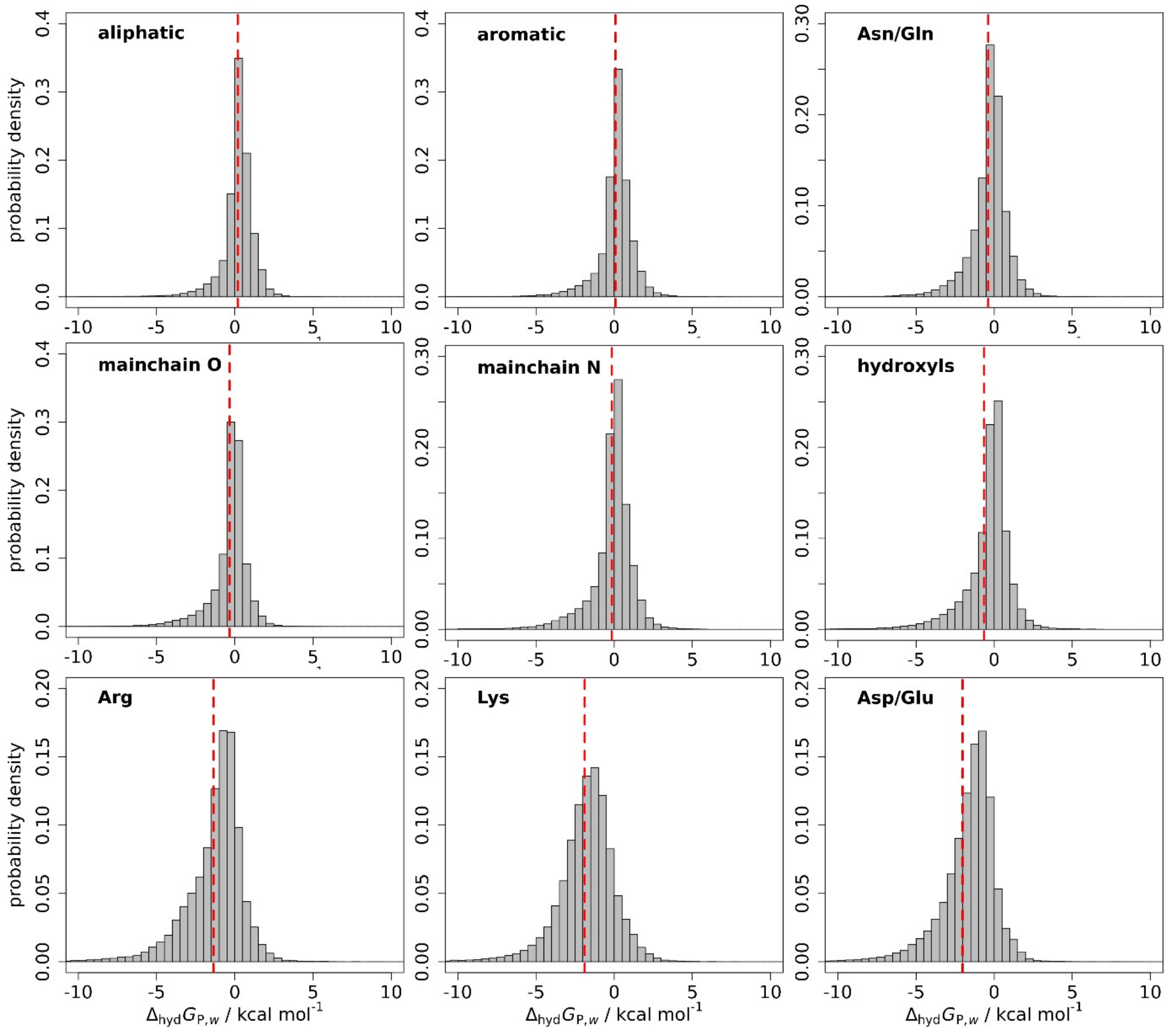

Figure 9: Probability densities of the calculated $\Delta_{hyd}G_P$ contributions (in kcal/mol) of water molecules in the proximity (3.5 Å) of specific protein groups (aliphatic atoms in Ala, Leu, Ile, Val, aromatic atoms in Trp, Phe, and Tyr, sidechain amides, mainchain N atoms, main chain O atoms, hydroxyls, carboxylates of Asp and Glu, guanidinium of Arg, amine of Lys) for complexes in the used PDBbind refined subset. All predicted water molecules were considered in this analysis. Respective average values are shown as red dashed lines. The probability density has the inverse unit of the x-axis parameter, i.e. kcal^{-1}·mol. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ data/ PDBbind_refined_set/ Figure9_Table9_Ghyd_AA/).

Despite the very intuitive overall trend, it has to be noted that standard deviations of the $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions are rather high, meaning that there are also water molecules near polar groups with unfavourable contributions and water molecules with favourable contributions near aliphatic sidechains. This shows that, although there are clear tendencies, water thermodynamics seem to be highly dependent on the precise properties and arrangement of the microenvironment. This underlines the need for accurate methods to elucidate the thermodynamic properties of specific binding site water molecules for drug design purposes.

4.1.1.6 Are replaced water molecules less "happy" than retained ones?

Another hypothesis that was investigated is whether the water molecules that are sterically displaced by ligand atoms have *per se* less favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions than the retained ones. This would massively facilitate the choice which water molecules should be targeted when designing a ligand.

Therefore, all predicted binding site water molecules in the used PDBbind refined subset were divided into two subsets, sterically replaced and retained water molecules. In this analysis, a binding site water molecule is defined as a water molecule within 3.5 Å of any ligand and protein atom, and "sterically replaced" denotes water molecules within a given threshold (here: 1.0 or 1.5 Å) of any ligand atom. The distribution of the respective $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions for the replaced and retained waters using a threshold of 1.0 and 1.5 Å are shown in Figure 10. In addition, the binding site water molecules were also divided into subsets based on their $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions, and the respective percentages of replaced water molecules within these subsets was determined (Table 10).

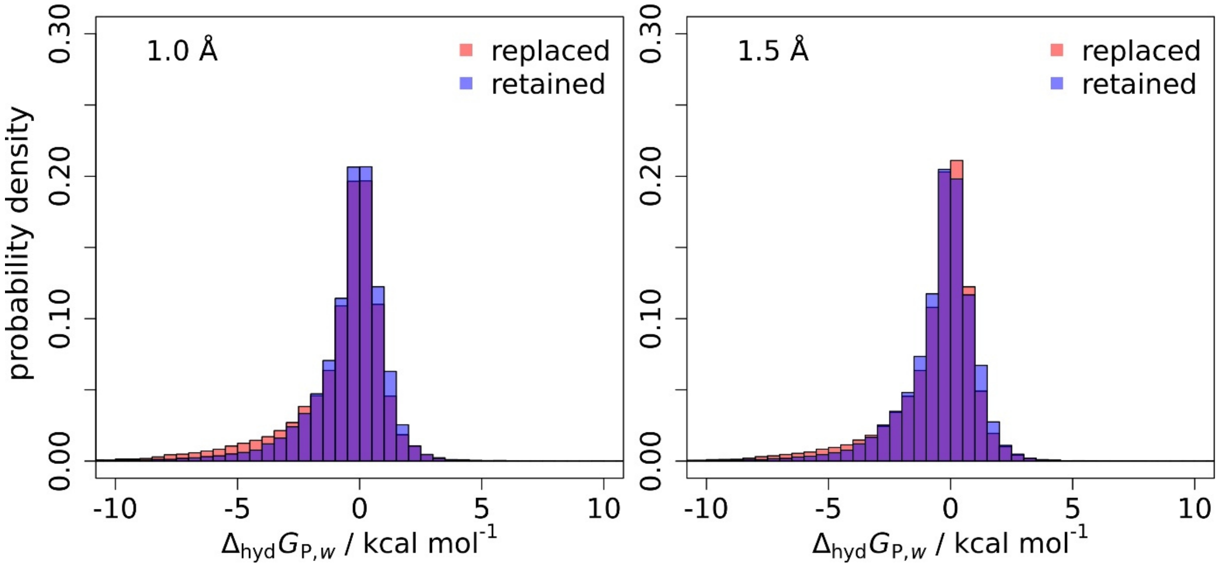

Figure 10: Probability densities of the $\Delta_{hyd}G_{P}$ contributions (in kcal/mol) for the replaced (red) and retained (blue) apo binding site water molecules within the used PDBbind refined subset as calculated by 3D RISM for replacement thresholds of 1.0 Å and 1.5 Å. A binding site water molecule denotes any water position within 3.5 Å of any ligand and protein atom. The probability density has the inverse unit of the x-axis parameter, i.e. kcal^{-1}·mol. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure10_Table10_replacement_Ghyd/).

Table 10: Percentages of replaced water molecules for subsets of binding site water molecules based on $\Delta_{hyd}G_P$ contributions (X % most and least favourable $\Delta_{hyd}\mathrm{G}_P$ contributions) for two replacement thresholds (1.0 and 1.5 Å) in the complexes of the used PDBbind refined subset. The respective percentage of retained water molecules is given implicitly since both values add up to 100 % (all water molecules are either replaced or retained). A binding site water molecule denotes any water position within 3.5 Å of any ligand and protein atom. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure10_Table10_replacement_Ghyd/).

	X % most favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions		X % least favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions	
	10 %	25 %	25 %	10 %
replaced (1.0 Å)	39.1 %	30.6 %	22.1 %	20.7 %
replaced (1.5 Å)	57.3 %	50.1 %	45.1 %	40.3 %

W.r.t. to the overall distribution of $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions within the replaced and retained subset (Figure 10), there are no significant differences, which is rather surprising since one would assume that "unhappy" water molecules should be replaced more easily, and hence more often, than "happy" water molecules (as replacement of a "happy" water molecule is associated with a respective penalty that has to be compensated by respective negative contributions to $\Delta_{\mathrm{bind}}H_{\mathrm{PL}}$).[192] Interestingly, when considering the percentages of replaced water molecules in the binding site water subsets with especially favourable and unfavourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions (Table 10), the trend is even reverse: the percentage of replaced water molecules is significantly higher for water molecules with the most favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions than for those with unfavourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions. This is an intriguing finding which leads to several questions and considerations. First, one has to keep in mind that the definition of the "replaced" set used here is an approximation; some of the "replaced" water molecules might actually be still present in the *holo* form but slightly shifted in their position. However, this would likely effect "happy" and "unhappy" waters alike and would not change the trends w.r.t. the different replacement percentages. Hence, the interesting question arises whether the reason for the higher replacement ratio within the "happy" water subset lies in physical principles or rather in a bias introduced by the design process of the ligands (in section 4.1.1.8, a detailed analysis correlating water "happiness" with ligand binding affinity will be presented). Polar ligand groups can compensate the enthalpic penalty of replacing a favourably bound water molecule by making favourable interactions with the respective microenvironment. As was shown in the microenvironment analysis, water molecules with favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions are preferentially located in a more polar environment with the possibility for hydrogen bonds. Such parts of a binding site are of course also highly attractive areas for the design of ligands since $\Delta_{\text{bind}}H_{\text{PL}}$ can be optimised via hydrogen bonds and ionic interactions. Water molecules with less favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions might be located in areas less attractive for design which might be a reason why they are less frequently replaced although this would result in a favourable contribution to $\Delta_{\text{bind}}G_{\text{PL}}$. However, it might also be the case that – for the ligands in the data set, which were often "designed" by mere trial and error – certain water molecules simply have to be replaced to achieve strong binding, or that factors like the formation of direct ligand-protein interactions outweigh the $\Delta_{\text{hyd}}G_{\text{P}}$ contribution by far. Yet, the presented findings are a reason to further investigate *apo* water thermodynamics since an analysis might reveal so far untargeted parts of a binding site where displacement of water molecules might not be obvious but highly favourable.

4.1.1.7 Influence of water thermodynamics on druggability

Previous studies[280,279,278] suggest a correlation between the thermodynamic properties of binding site water molecules and the respective druggability of the protein binding site. Some of the complexes in the used PDBbind refined subset are part of the NRDLD set by Krasowksi *et al.*[280] which provides a classification into druggable and undruggable structures (list in Appendix, 7.3). To evaluate if the trend observed in literature can also be seen in the present study, the $\Delta_{\text{hyd}}G_{\text{P}}$ contributions of the predicted binding site water molecules within the druggable and undruggable structure subsets were investigated. The distribution of $\Delta_{\text{hyd}}G_{\text{P}}$ contributions in the druggable and undruggable set are shown in Figure 11. In addition, the average $\Delta_{\text{hyd}}G_{\text{P}}$ contributions and respective percentile values for the 10 % and 25 % of water molecules with the most and least favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions are given in Table 11.

In accordance with literature, the results show that the average $\Delta_{\text{hyd}}G_{\text{P}}$ contributions of water molecules in undruggable binding sites are much more favourable than for druggable binding sites. In average, the 25 % "unhappiest" water molecules in undruggable binding sites have $\Delta_{\text{hyd}}G_{\text{P}}$ contributions only slightly larger than for the 25 % "happiest" water molecules in druggable binding sites.

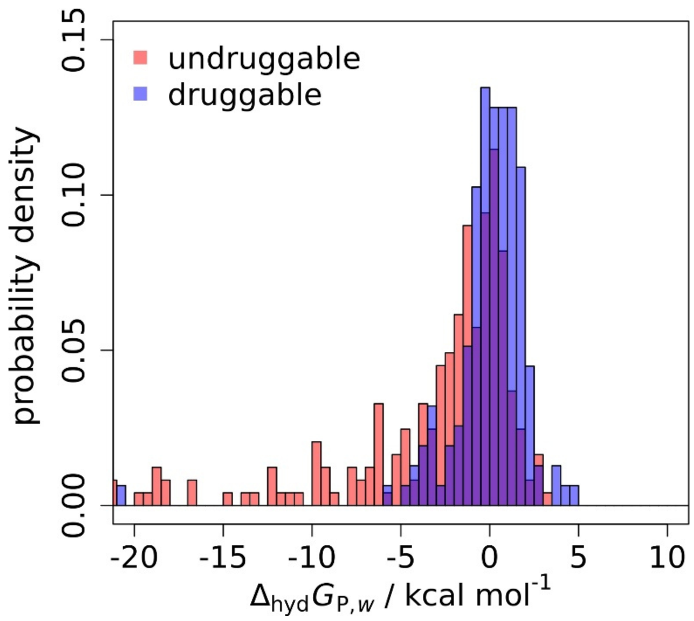

Figure 11: Probability densities of the $\Delta_{hyd}G_P$ contributions (in kcal/mol) for the binding site water molecules in the structures of the druggable (blue) and undruggable (red) protein subset. A binding site water molecule is defined as any predicted water position within 3.5 Å of any protein and ligand atom. The druggable and undruggable subsets are defined as the intersection between the used PDBbind refined subset and the NRDLD set (PDB codes given in SI 7.3). The probability density has the inverse unit of the x-axis parameter, i.e. kcal-1·mol. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure11_Table11_ Ghyd_druggability/).

Table 11: Average $\Delta_{hyd}G_P$ contributions (in kcal/mol) of binding site water molecules in the structures of the druggable and undruggable protein subset as well as respective $\Delta_{hyd}G_P$ percentile thresholds for the water molecules with the X % most and least favourable $\Delta_{hyd}G_P$ contributions. A binding site water molecule is defined as any predicted water position within 3.5 Å of any protein and ligand atom. The druggable and undruggable subsets are defined as the intersection between the used PDBbind refined subset and the NRDLD set (PDB codes given in Appendix, 7.3). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ data/ PDBbind_refined_set/ Figure11_Table11_ Ghyd_druggability/).

		$\Delta_{hyd}G_P$ contributions / kcal·mol^{-1}			
		percentile thr. X % most favourable contributions		percentile thr. X % least favourable contributions	
	mean	10 %	25 %	25 %	10 %
druggable	0.01 ± 2.41	-2.14	-0.71	1.37	1.92
undruggable	-3.50 ± 6.17	-11.13	-4.57	0.15	0.91

This is especially interesting when considering the results from the replaced/retained analysis: The finding that undruggable binding sites contain more “happy” water molecules suggests that one reason for their undruggability is that the water molecules there are tightly bound and thus more difficult to replace. At first glance, this seems to be somewhat contradictory to the finding that, when considering the whole data set, the ratio of replaced water molecules is even slightly higher for those hydration sites with especially favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions. This seeming discrepancy, however, could result from different effects: As already indicated, rather “unhappy” water molecules are maybe less often replaced than “happy” ones not because they are tightly bound and hard to replace but because they simply have not been targeted during the design process. Furthermore, the presented replaced/retained analysis is a descriptive study of the available data, but it does *per se* not provide any assessment if the found preferences are “optimal” w.r.t. ligand binding. Such an assessment is difficult to achieve since it would in principle require knowledge about the “optimal” ligand for any given binding site. W.r.t. to the given data set, however, a step towards such an evaluation is done in the next section by taking into consideration the provided affinity data.

4.1.1.8 Influence of water thermodynamics on ligand affinity

The idea is that, if the replacement of “unhappy” water molecules is on average more favourable for $\Delta_{\mathrm{bind}}G_{\mathrm{PL}}$ than replacement of a “happy” water molecule, the average properties of the replaced *apo* water molecules should be different for ligands with very high and very low affinity (although water replacement is of course only one aspect that contributes to $\Delta_{\mathrm{bind}}G_{\mathrm{PL}}$).

Therefore, the used PDBbind refined subset was divided into subsets based on ligand affinity (structures with the 1 %, 5, %, 10 %, and 25 % most and least affine ligands), and the average $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions of the respective binding site water molecules were analysed. In Table 12, the respective average $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions of i) all (= replaced + retained) binding site water molecules, ii) the replaced binding site water molecules, and iii) the retained binding site water molecules are listed for all affinity subsets. The illustrative distribution of the $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions of the replaced water molecules for complexes with the 5 % most and least affine ligands is shown in Figure 12.

Table 12: Average calculated $\Delta_{hyd}G_P$ contributions (in kcal/mol) for all, replaced, and retained predicted apo binding site water molecules for subsets of complexes in the used PDBbind refined subset based on the corresponding affinity of the ligand (eight bins corresponding to complex structures with the 1 %, 5 %, 10 %, and 25 % most and least affine ligands. Respective pK_i/pK_d thresholds (combined as $pK_{aff.}$) are given in column 1; as explained in 2.1, K_i denotes the special case of an equilibrium constant for the dissociation process of an inhibitor-enzyme complex, so that both values are mixed within the PDBbind refined set. For the replaced/retained separation, a threshold of 1.0 Å was used. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure12_Table12_Ghyd_affinity/).

affinity subset	av. $\Delta_{\text{hyd}}G_{\text{P}}$ contributions / kcal·mol^{-1}		
	all	replaced (1.0 Å)	retained (1.0 Å)
Max 1 % (p$K_{\text{aff.}}$ > 11.2)	-0.46 ± 2.77	-0.89 ± 3.48	-0.32 ± 2.47
Max 5 % (p$K_{\text{aff.}}$ > 9.7)	-0.71 ± 4.41	-1.38 ± 5.96	-0.48 ± 3.73
Max 10 % (p$K_{\text{aff.}}$ > 9.0)	-0.71 ± 4.40	-1.51 ± 6.44	-0.43 ± 3.38
Max 25 % (p$K_{\text{aff.}}$ > 7.9)	-0.79 ± 4.58	-1.71 ± 6.92	-0.47 ± 3.36
Min 25 % (p$K_{\text{aff.}}$ < 5.0)	-1.19 ± 5.59	-1.92 ± 7.54	-0.96 ± 4.80
Min 10 % (p$K_{\text{aff.}}$ < 3.8)	-1.28 ± 6.45	-2.07 ± 8.51	-1.04 ± 5.68
Min 5 % (p$K_{\text{aff.}}$ < 3.2)	-1.51 ± 5.99	-2.67 ± 8.61	-1.17 ± 4.92
Min 1 % (p$K_{\text{aff.}}$ < 2.3)	-1.74 ± 6.38	-3.61 ± 10.43	-1.18 ± 4.32

Intriguingly, the analysis indeed reveals significant differences for the high and low affinity structures. W.r.t. all binding site water molecules, it can be seen that the average $\Delta_{\text{hyd}}G_{\text{P}}$ contributions are slightly higher (and thus less favourable) for structures with the most affine ligands (i.e. the highest pK_{d}/pK_{i} values). Further separation of these binding site water molecules into replaced and retained ones reveals a strong trend for the replaced waters and only a slight trend for the retained ones: As can also be seen in Figure 12, in average, the water molecules which get replaced by highly affine ligands are significantly less "happy" than those that get replaced by the least affine ligands. For the retained water molecules, much smaller differences are observed. This is a highly relevant finding since the stronger trend for the replaced water molecules compared to the retained ones implies that replacement of "unhappy" water molecules in the binding site indeed correlates with a higher binding affinity.

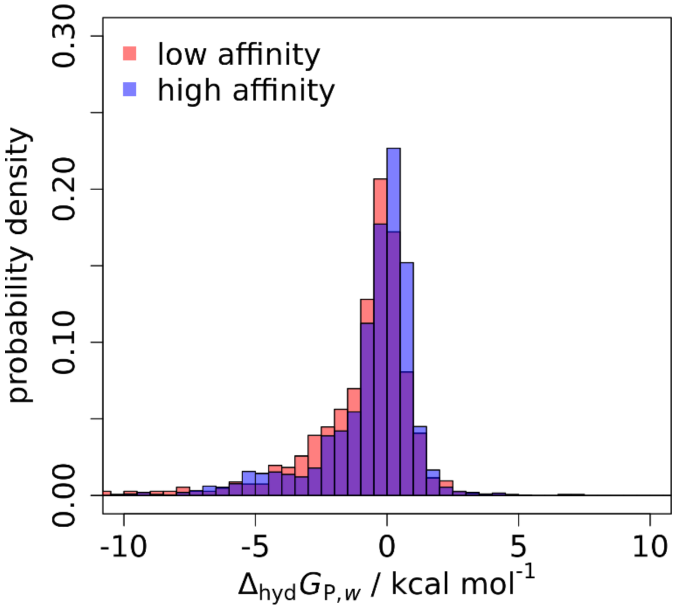

Figure 12: Probability densities of the calculated $\Delta_{hyd}G_P$ contributions (in kcal/mol) for the predicted replaced binding site water molecules in structures of complexes with the 10 % most (blue) and least (red) affine ligands within the used PDBbind refined subset. The probability density has the inverse unit of the x-axis parameter, i.e. kcal^{-1}·mol. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure12_Table12_Ghyd_affinity/).

Following this conclusion, the high replacement ratios of "happy" water molecules for the overall data set might be interpreted – as least to a certain extent - as result of a non-optimal ligand design. An unambiguous assessment however is difficult to achieve in this case. As already mentioned in 4.1.1.6, many of the investigated ligands were "designed" via trial and error, and it is hardly possible to disprove the hypothesis that certain water molecules simply had to be replaced (independent of their "happiness") to allow for necessary protein-ligand interactions. Yet, the results of this study suggest that knowledge about the thermodynamic properties of binding site water molecules could indeed be exploited for optimizing ligand affinity. As a consequence, the results also highlight the need for the rigorous assessment of trends extracted from large data sets. Especially with the re-emergence of ML methods, which make excessive use of available structural data and ligand chemistry, methods are needed to critically assess if commonly observed trends in the data can be attributed to physical principles or rather to common but maybe non-optimal design principles. Approaches like the one presented in this work can thus help to improve future drug design strategies.

4.1.1.9 Water displacement by ligand atoms - replacement propensities

As a first step towards water replacement rules for drug design, it was investigated if atoms of a certain element type (hydrogen, carbon, nitrogen, oxygen, phosphorous, sulphur, fluorine, and chlorine) displace water molecules more frequently than others.

To develop a quantitative measure for the replacement propensities of the respective elements at different distances to a water position, a data set-wide pair distribution function of each element w.r.t. the oxygen water position can be analysed. Such a pair distribution function shows the propensity of finding the selected element at any given distance to a water oxygen atom.

Here, the continuous pair distribution function is approximated by calculating the replacement propensity $p(X_i,d)$, i.e the probability of finding an atom of a given element X_i within a given distance bin d (here: 0.1 Å bins from 0.0 to 3.0 Å) to a predicted water position in relation to all other elements, based on all complex structures in the used PDBbind refined subset, according to (69):

$$p(X_i,d) = \frac{N(X_i,d)}{\sum_{j=1}^{X_N} N(X_j,d)} \tag{69}$$

Here, $N(X_i,d)$ denotes the number of atoms of element X_i that are found within a distance bin d of any predicted water molecule within the PDBbind refined subset. For instance, a value of 0.53 for oxygen in the 0.1 – 0.2 Å bin means that 53 % of all investigated atoms that are found within a distance of 0.1 to 0.2 Å of any predicted water position in the whole PDBbind refined subset are oxygen atoms (and thus that all other elements combined only make up 47 %).

However, to get a complete picture, one should also consider the relative frequency of a given element among all ligand atoms in the PDBbind refined subset since e.g. hydrogen and carbon atoms are much more abundant than fluorine atoms and thus naturally will replace water molecules more frequently. To account for this, also a normalised displacement propensity value $p^*(X_i,d)$ was determined which is normalised by the relative occurrence of the given element among all investigated elements according to (70):

$$p^*(X_i,d) = \frac{p(X_i,d)}{N(X_i) / \sum_{j=1}^{X_N} N(X_j)} \tag{70}$$

Here, $N(X_i)$ denotes the total number of atoms of element type X_i in all ligands in the PDBbind refined subset. Consequently, the $p^*(X_i,d)$ shows if atoms of a certain element type replace water molecules more or less often than would be expected from their occurrence: If $p^*(X_i,d) > 1$, it is overrepresented among atoms that replace water molecules at a given distance, if $p^*(X_i,d) < 1$, it is underrepresented.

Both the *p(X_i,d)* and *p^*(X_i,d)* values for hydrogen, carbon, nitrogen, oxygen, phosphorous, sulphur, fluorine, and chlorine atoms for distances between 0.0 and 3.0 Å are plotted in Figure 13 (left: $p(X_i,d)$, right: normalised $p^*(X_i,d)$).

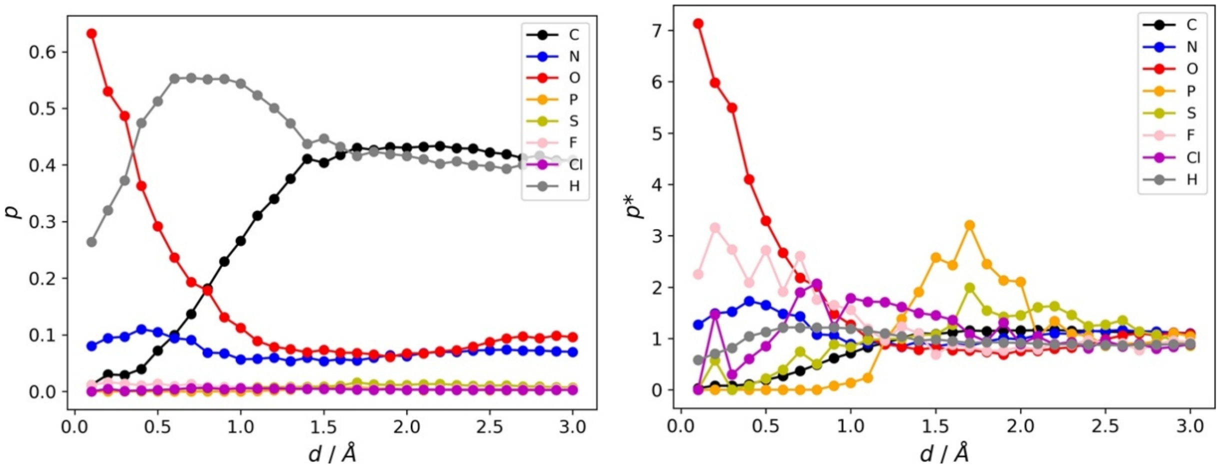

Figure 13: Unnormalised and normalised displacement propensities $p(X_i,d)$ (left panel) and $p^*(X_i,d)$ (right panel) for hydrogen, carbon, nitrogen, oxygen, phosphorous, sulphur, fluorine, and chlorine as defined in Eq. (69) and (70) for the used PDBbind refined subset for distance bins d between 0.0 and 3.0 Å in 0.1 Å increments. Unnormalised $p(X_i,d)$ values denote the fraction of atoms of element type X_i among all regarded atoms present at a given distance to any water molecule within the data set, normalised $p^*(X_i,d)$ values indicate if atoms of element type X_i are over- or underrepresented among all regarded atoms present at a given distance w.r.t. to their natural occurrence. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ data/ PDBbind_refined_set/ Figure13_displacement_propensities/).

When looking at the unnormalised displacement propensities in Figure 13, the by far highest values at small distances can be observed for O atoms. This is in accordance with other studies[76] and highly intuitive since a ligand oxygen atom is well suited to mimic the water oxygen role and maintain the respective interactions, especially hydrogen bonds, with the water molecule's microenvironment. When taking into account the relative frequency of the elements (Figure 13, right panel), it can be seen that, for distances up to 0.5 Å, oxygen atoms can be found three to seven times more often than would be expected for a uniform probability among the investigated elements.

By absolute values, hydrogen atoms are the second most abundant element in the proximity of *apo* water positions, with a peak at distances between 0.7 and 1.0 Å. This can be attributed to the fact that hydrogen atoms are bound to many replacing heavy atoms. Due to its high frequency, only values around 1 are obtained for the normalised replacement propensities, indicating no enrichment among the water replacing atoms.

A similar trend is observed for carbon atoms: They are only rarely found to displace water molecules directly although they are by far the most frequent element type. For small distances, they are clearly underrepresented with $p^{*}(\text{carbon},d) < 1$, and for distances > 1 Å, the displacement propensities roughly correspond to the expected value based on frequency with $p^{*}(\text{carbon},d) \approx 1$.

The second most frequent heavy atom type to replace water molecules are nitrogen atoms, which are rather polar and potentially capable of reproducing the water molecule's interactions. When taking into account the overall frequency, nitrogen is slightly overrepresented at distances around 0.5 Å.

Interestingly, a different tendency is observed for fluorine, phosphorous, sulphur, and chlorine: While these elements have low unnormalised displacement propensities because of their low overall frequency, there are clear trends for the normalised displacement propensities: Phosphorous is overrepresented at distances of around 1.3 Å to 2.0 Å. This is likely an indirect effect since phosphorous is present in the ligands within phosphate groups whose oxygen atoms are likely to replace the water molecules. A similar trend is observed for sulphur which occurs in thiol groups, but also in oxygen-containing sulfone or sulfoxide groups. Fluorine atoms are overrepresented among the replacing ligand atoms below 1.0 Å, and chlorine between 0.7 Å and 1.5 Å, even exceeding the relative probabilities of nitrogen. This is interesting since halogen atoms have rather different properties than a water oxygen atom und should not be able to retain its hydrogen bonds.

This directly leads to the question if displacement probabilities are different for water molecules with different thermodynamic properties. Therefore, the normalised displacement propensities were determined for two subsets of water molecules with different properties, namely the 25 % of predicted water molecules in the used PDBbind refined subset with the most and least favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions. The respective results are given in Figure 14.

For oxygen atoms, no large difference can be seen w.r.t. the two subsets: Both "happy" and "unhappy" water molecules alike are frequently replaced by oxygen-containing groups. This suggests that the respective replacement is dominated by enthalpic effects, i.e. the formation of favourable interactions between the ligand and the binding site which compensate or outweigh potential penalties w.r.t. solvation effects.[192] A completely different picture, however, is obtained for the other element types: Nitrogen atoms are clearly overrepresented at small distances to "happy" water molecules but not in the proximity of "unhappy" water molecules. An even more pronounced effect can be observed for phosphorous and sulphur: Both elements are highly overrepresented at approx. 1.5 Å distance to water molecules with the most favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions, with values exceeding those of oxygen, but show values around 1 for the water molecules with the least favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions. This suggests that especially “happy” water molecules tend to get replaced by ligand phosphate groups and S-containing groups like thiols or sulfones. These functional groups are charged or highly polar, being in accordance with the finding that “happier” water molecules tend to be located in more polar regions of the protein. Concordantly, “happy” water molecules are practically never displaced by fluorine or chlorine atoms which are rather nonpolar and not able to mimic a water molecule’s polar interactions. Both elements are, however, highly overrepresented among the atoms displacing “unhappy” water molecules: For distances up to 1.5 Å, they exhibit normalised replacement propensities partially exceeding those of oxygen. This suggests, that here – opposed to the oxygen atoms -, the replacement is dominated by solvation effects rather than by enthalpic contributions. This is an important finding since it allows to derive direct rules for drug design based on the thermodynamic signature of the *apo* binding site water molecules.

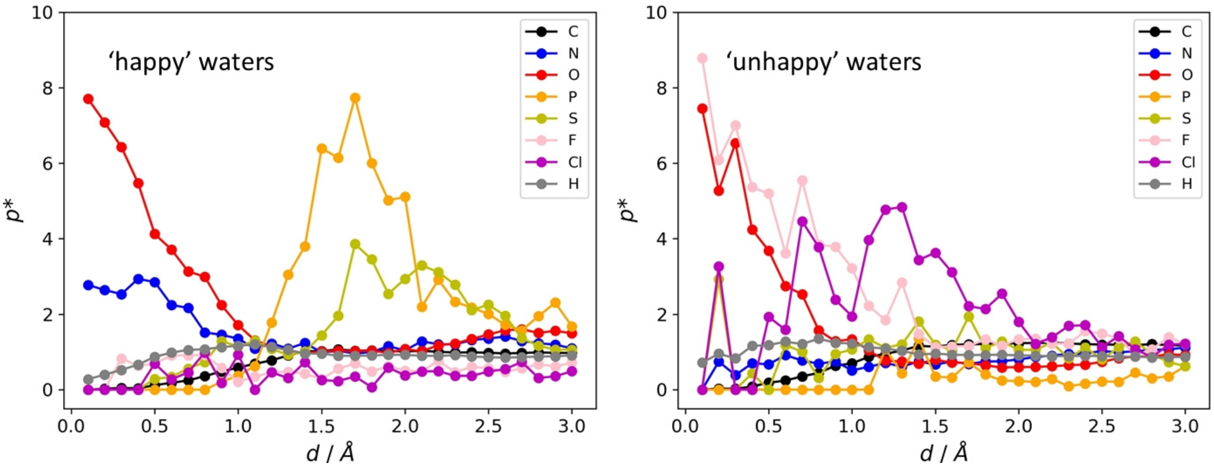

Figure 14: Normalised displacement propensities $p^*(X_i,d)$ for hydrogen, carbon, nitrogen, oxygen, phosphorous, sulphur, fluorine, and chlorine as defined in Eq. (70) for two water subsets containing only those water molecules within the used PDBbind refined subset that exhibit the 25 % most favourable (panel a, “happy”) and 25 % least favourable (panel b, “unhappy”) $\Delta_{hyd}G_P$ contributions, with distance bins d between 0.0 and 3.0 Å in 0.1 Å increments. The normalised $p^*(X_i,d)$ values indicate if atoms of element type X_i are over- or underrepresented among all regarded atoms present at a given distance w.r.t. to their natural occurrence. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ data/ PDBbind_refined_set/ Figure14_displacement_propensities_happy_unhappy/).

4.1.1.10 Mapping water thermodynamics onto ligand atoms

To elucidate further if water molecules with certain properties are preferentially replaced by certain atoms, it is highly desirable to directly map *apo* water thermodynamics onto specific ligand atoms. As outlined in the computational details section, the calculations and algorithms used in this work allow for a fast and efficient way to achieve this: After applying a Gaussian convolution on the free energy density field, the resulting field can simply be evaluated at a respective ligand atom position to obtain an approximation of the $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contribution of the *apo* hydration site that was replaced by this ligand atom.

An illustrative example is shown in Figure 15 for a fXa complex structure. Overlay of the predicted *apo* water positions, coloured according to their $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions, and the *holo* ligand, coloured by the interpolated $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions on each atom *l*, shows that the interpolation nicely captures the *apo* water thermodynamics w.r.t. replacing ligand groups. For instance, it directly reveals the replacement of a highly unstable hydration site by the chlorine substituent, whose effect on SAR was already investigated in earlier work.[192] With the interpolated $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions, there is now a tool at hand to directly correlate ligand chemistry with *apo* water thermodynamics in an efficient manner.

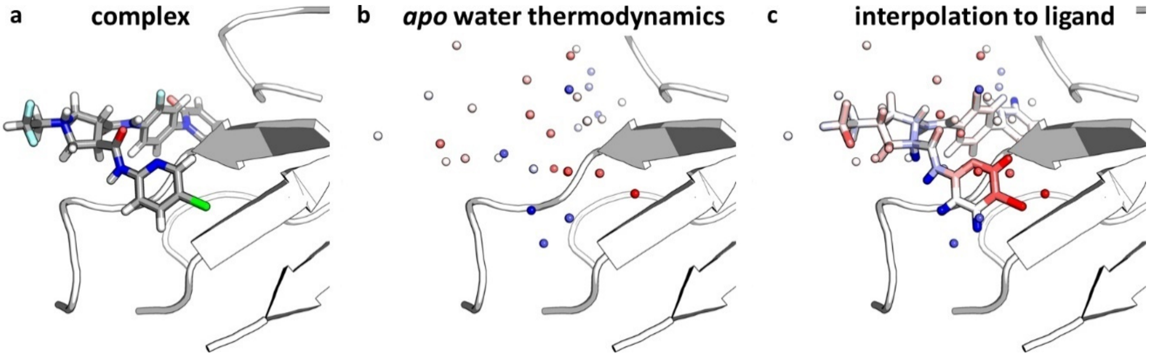

Figure 15: Illustration of the interpolation of apo water thermodynamics to ligand atoms; a) complex 2xbv[276]; b: apo water molecules (within 4.0 Å of any ligand atom) as predicted by 3D RISM-based algorithms, coloured by their calculated $\Delta_{hyd}G_P$ contributions (from blue to red from -2.0 to +2.0 in units of kcal/mol); c: overlay with the ligand, coloured by the interpolated $\Delta_{hyd}G_P$ contributions (same colour code). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ structures/ 2xbv/).

For a set of various GAFF atom types, the average interpolated *apo* water $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions on all respective ligand atoms *l* in the used PDBbind refined subset were determined. However, the interpolated value $\Delta_{\mathrm{hyd}}G_{\mathrm{P},l}$ alone does not capture if a specific ligand atom does actually replace a distinct hydration site. To include this information, only those ligand atoms with high interpolated $g_{\mathrm{O}}(\mathbf{r})$ values (here: > 3; other thresholds were also tested but did not yield different trends) were considered in the analysis, i.e. only atoms which coincide with a predicted water position and can thus be considered a replacing ligand atom. The respective average $\Delta_{\mathrm{hyd}}G_{\mathrm{P},l}$ values of atoms of the different GAFF atom types for atoms with $g_{\mathrm{O}}(\mathbf{r})$ values > 3 are given in Table 13 in the first column (“all”).

When interpreting the results, it is useful to keep in mind that there are – as worked out in earlier work by the Kast group[192] – basically two cases w.r.t. the replacement of an apo water molecule upon ligand binding: If a hydration site with a positive, hence unfavourable, $\Delta_{\text{hyd}}G_{\text{P}}$ contribution is replaced, the solvation contribution to $\Delta_{\text{bind}}G_{\text{PL}}$ is favourable. If a water molecule with negative, hence favourable, $\Delta_{\text{hyd}}G_{\text{P}}$ contribution is replaced, this penalty must be compensated by respective, favourable contributions to $\Delta_{\text{bind}}H_{\text{PL}}$, i.e. by formation of interactions between the ligand and the binding site residues. Since ionic and polar interactions (salt bridges, hydrogen bonds) are stronger than van der Waals interactions, it is to be expected that ionic and polar ligand groups can replaced "happy" und "unhappy" waters alike (since their interactions should be strong enough to compensate the associated penalty), while apolar groups likely replace more "unhappy" water molecules.

Table 13: Average interpolated $\Delta_{hyd}G_P$ contributions (in kcal/mol) of the replaced apo water molecules at the respective ligand atom positions for ligand atoms of specific GAFF atom types (n4, n, o, oh, c3, ca, f, cl) in the used PDBbind refined subset that exhibit interpolated g_O values > 3. Results are given for i) all ligand atoms of a given type that fulfil the $g_O > 3$ criterion (all), ii) for respective ligand atoms belonging to the 25 % most affine ligands (max 25 %), and iii) for respective atoms belonging to the 25 % least affine ligands (min 25 %) within the used PDBbind refined subset. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure16_Table13_ Ghyd_atomtypes/).

Atom type	av. interpolated $\Delta_{\text{hyd}}G_{\text{P}}$ contributions / kcal·mol^{-1}		
	all	Max 25 % (p$K_{\text{aff.}}$ > 7.9)	Min 25 % (p$K_{\text{aff.}}$ < 5.0)
n4 (N w. four substituents)	-3.47 ± 2.44	-3.16 ± 2.04	-3.34 ± 2.77
n (N in amides)	-1.82 ± 5.46	-0.99 ± 1.44	-3.73 ± 9.41
o (O in carbonyl or carboxyl)	-3.15 ± 7.69	-2.18 ± 5.86	-4.31 ± 9.95
oh (O in hydroxyl)	-2.04 ± 5.85	-2.63 ± 4.69	-1.16 ± 3.46
c3 (sp3 C)	-0.58 ± 1.43	-0.34 ± 1.02	-0.85 ± 2.05
ca (aromatic carbon)	-0.14 ± 0.91	-0.06 ± 0.79	-0.11 ± 0.92
f (any F)	0.15 ± 1.07	0.12 ± 1.24	-0.02 ± 1.28
cl (any Cl)	0.38 ± 1.12	0.75 ± 1.05	0.22 ± 0.99

Indeed, the results reveal large differences between the selected GAFF atom types. In average, halogen atoms displace water molecules with the least favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions. This is in line with the analysis of the replacement propensities, which shows a clear overrepresentation of fluorine and chlorine atoms in the proximity of "unhappy" water molecules. They also exhibit rather low standard deviations, supporting the hypothesis that displacement can only be observed for those "unhappy" water molecules since the penalty for displacing a water molecule with favourable interactions cannot be compensated by favourable contributions to $\Delta_{\mathrm{bind}}H_{\mathrm{PL}}$ via i.e. hydrogen bonds or salt bridges. A similar picture is obtained for carbon atoms (GAFF types c3 and ca), which were shown to generally be underrepresented within water replacing ligand atoms. If they replace a water molecule, then rather those with relatively unfavourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions.

Charged or polar ligand atoms, on the other hand, tend to replace water molecules with more favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions in average. This trend is especially pronounced for n4, a positively charged nitrogen, which seems to replace exclusively "happy" water molecules. For the other polar or charged atom types, the average $\Delta_{\mathrm{hyd}}G_{\mathrm{P},l}$ values are also negative, however with high standard deviations, implying that especially atoms of type o and oh also replace a lot of water molecules with unfavourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions. Similar to the replacement analysis, this finding again poses an interesting question: To what extent are the displacement preferences observed in the used PDBbind refined subset based on natural principles, and to what extent does one rather measure the effect of the design process? For instance, the displacement of "unhappy" water molecules by polar groups could be an indirect effect of a medicinal chemist's intuition to displace a crystallographic water with an oxygen atom.

As already outlined, dissecting the found statistics into general principles and bias is a difficult task since there is no knowledge about the "perfect" ligand for each structure, which would in principle allow to draw unambiguous conclusions. However, again an attempt was made towards such an assessment by taking into consideration the affinity values of the given ligands. Following the assumption that the most affine ligands within the data set are the most optimal w.r.t. to improving $\Delta_{\mathrm{bind}}G_{\mathrm{PL}}$, the average $\Delta_{\mathrm{hyd}}G_{\mathrm{P},l}$ values of the selected GAFF atom types were determined considering only atoms belonging to the 25 % most and least affine ligands in the used PDBbind refined subset. The results are given in Table 13 in column two and three, and the respective histograms are shown in Figure 16.

Comparison of the average $\Delta_{\mathrm{hyd}}G_{\mathrm{P},l}$ values for the whole data set and for subsets with the most and least affine ligands reveals some striking trends: In accordance with the replacement analysis, values for the most affine ligands are generally shifted to less negative values as it was shown that the most affine ligands generally replace more "unhappy" water molecules. This could in principle be interpreted as a kind of "evolutionary selection" during the ligand design process: Even if factors like water replacement are not directly included in the design process, the best ligands w.r.t. binding affinity get selected and further optimised over several stages in the drug design process, so that – in many cases - those which are most suitable for the given binding site (including solvation effects) naturally prevail.

Consequently, the preference for the replacement of “unhappy” water molecules by carbon and especially halogen atoms is even more pronounced for the most affine ligands, albeit the differences are rather small.

For the n4 atom type, the results for both subsets are similar to those for the overall set, with the n4 atoms replacing exclusively “happy” water molecules (Figure 16). For the other polar atom types, however, highly interesting trends can be observed: In the most affine ligands, the average $\Delta_{\text{hyd}}G_{\text{P},l}$ value for the atom type n is much higher than for the overall set and for the least affine ligands, with rather small standard deviations, as can also be seen in the respective histogram (Figure 16). Given the fact that the differences for the atom types c3, ca, f, cl, and n4 between the subsets are rather small, with a small shift to more unfavourable $\Delta_{\text{hyd}}G_{\text{P},l}$ values for most affine ligands, the massive differences observed for the n atom type can be considered meaningful. The trend suggests that the replacement of especially “happy” water molecules with amide groups correlates with lower ligand affinity, indicating that an amide group might not be ideally suited to compensate the energetic penalty associated with this replacement. Hence, some ligands might be improved w.r.t. affinity when substituting an amide group with for instance a hydroxyl group.

Intriguingly, a similar finding is observed for the o atom type in carbonyl and carboxyl groups: In the least affine ligands, respective atoms replace much more “happy” water molecules (Figure 16) than those in the most affine ligands. This is a surprising and important finding since oxygen atoms are the atoms which replace by far the most water molecules, and adding a carbonyl or carboxyl group to target a crystallographic water for replacement seems highly intuitive. The trend observed here, however, suggests that this might not always be an optimal strategy.

The most striking and important difference, however, can be seen of the oh atom type: In contrast to the overall observed shift to less negative $\Delta_{\text{hyd}}G_{\text{P},l}$ values for most affine ligands, the average $\Delta_{\text{hyd}}G_{\text{P},l}$ value of oh atoms within the most affine ligands is much more negative than for the least affine ligands and for the whole data set, as can be also seen in the histogram (Figure 16). This allows to draw conclusions that are highly relevant for drug design: Obviously, in the least affine ligands, many “unhappy” water molecules get replaced by hydroxyl groups. This is likely due to the general design strategy of replacing a crystallographic water molecule with a similar functional group in the ligand. However, the results obtained here imply that this is a suboptimal choice w.r.t. binding affinity when the respective water is an “unhappy” one. Likely, the microenvironment which leads to highly unfavourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions of the water molecule is likewise not ideal for accommodating a hydroxyl group. Here, rather substitution of the hydroxyl group with a carbon or halogen atom would be beneficial for binding affinity. This observation thus illustrates that the overall observed trend - that ligand atoms of type o and oh replace water molecules with a wide range of thermodynamic properties – is indeed not based on thermodynamic principles but rather an effect imposed by the common design strategy to replace crystallographic water molecules with oxygen-containing ligand groups. However, in many cases, the "evolutionary selection" discussed above naturally led to selection of a highly suited ligand, generally resulting in rather slight trends with high fluctuations.

Yet, by revealing this bias, important conclusions can be drawn for new design principles: Following the presented analysis, hydroxyl group should be used to replace especially "happy" water molecules since they can best take up the water molecule's interactions. Carbonyl groups, on the other hand, do not seem to be a good choice for replacing especially "happy" water molecules, probably because they are not that well suited to mimic the water molecule's hydrogen bonds. The same holds true for amide groups. "Unhappy" water molecules, on the other hand, should be targeted by aromatic or aliphatic groups and especially by halogen atoms, but can also be replaced by more polar groups. Here, the microenvironment and the possibilities for hydrogen bonds are likely the determining factor since replacement of an "unhappy" water should generally be beneficial for $\Delta_{\mathrm{bind}}G_{\mathrm{PL}}$ but favourable interactions lead to further improvement via contributions to $\Delta_{\mathrm{bind}}H_{\mathrm{PL}}$. This study thus confirms the principles presented in earlier work[192] in a large-scale analysis und can help to optimise future drug design processes w.r.t. solvation effects.

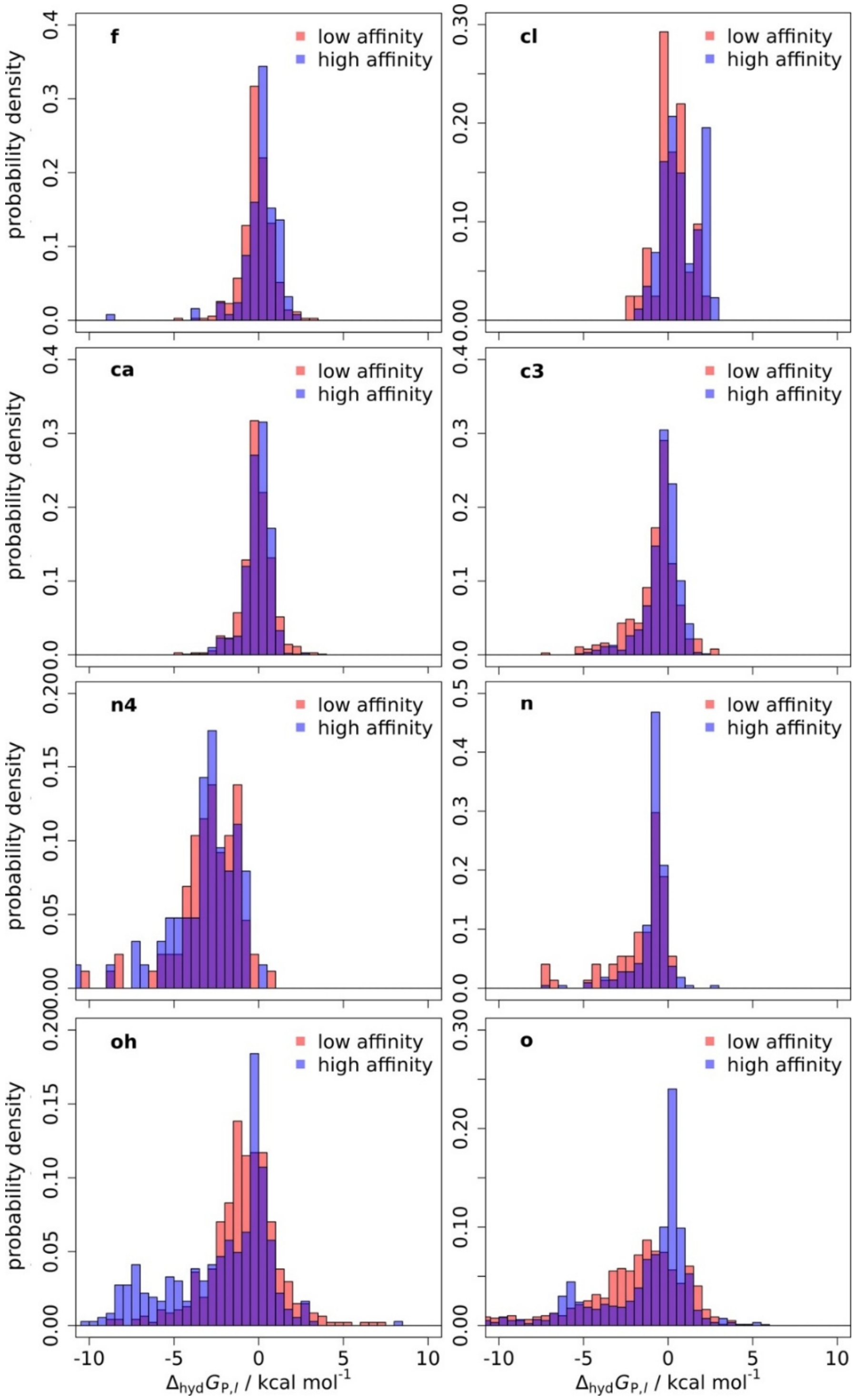

Figure 16: Probability densities of interpolated $\Delta_{hyd}G_P$ contributions (in kcal/mol) of the displaced apo water molecules at the respective atom positions for ligand atoms of specific GAFF atom types (f, cl, ca, c3, n4, n, oh, o) belonging to the 25 % most affine (blue) and least affine (red) ligands in the used PDBbind refined subset. Only ligand atoms that exhibit interpolated g_O values > 3 were considered in the analysis. The probability density has the inverse unit of the x-axis parameter, i.e. kcal^{-1}·mol. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ data/ PDBbind_refined_set/ Figure16_Table13_ Ghyd_atomtypes/).

4.1.1.11 Summary

In the presented study, novel physics-based approaches derived from 3D RISM integral equation theory were used to determine the local thermodynamic properties of water molecules within binding sites of a large data set containing more than 3800 protein structures.

Correlation of the respective results with experimental data and ligand chemistry allowed to gain a complete picture of the characteristics and the relevance of protein hydration sites: It was shown that water "happiness" is determined by the precise microenvironment within the binding site, and that highly localised water molecules with low experimental B-factors are not *per se* more "happy" than other water molecules.

W.r.t. to the replacement of water molecules, the analysis revealed that, although "happy" and "unhappy" water molecules are replaced alike within the whole data set, replacement of more "unhappy" hydration sites favourably correlates with ligand binding affinity and druggability. Following this conclusion, replacement preferences for different elements and atom types were investigated to derive practical rules for drug design. By correlating the found trends with provided binding affinity data, a hidden bias in the data set could be identified. For instance, it was shown that the replacement of a water molecule with an oxygen atom is not always optimal but rather depends on the specific thermodynamic properties of the targeted hydration site. Based on this analysis, improved replacement strategies were introduced that can now be employed for the rational design and optimisation of ligands. Thus, the presented study also showed that some trends derived from large-scale data might actually be biased by common design principles in medicinal chemistry and thus highlights the need for the critical assessment of available data to gain an optimal benefit for future research.

4.1.2 Analysis of *holo* water thermodynamics

Chapter 4.1 dealt with the characteristics of *apo* protein hydration sites and their correlation with ligand features. However, especially w.r.t. the optimisation of an existing ligand, it can also be of interest to characterise the water molecules in the proximity of a bound ligand. Such an analysis might for instance reveal newly introduced unstable hydration sites which could be replaced by addition of further substituents. Therefore, in this chapter, the thermodynamic properties of *holo* water molecules will be analysed and compared with those from the *apo* binding sites.

4.1.2.1 Reproduction of experimental water positions

As in 4.1, a prerequisite for any further analysis is the correct placement of the water molecules. In this part of the work, an even better matching between predicted and experimental water position is expected: While in 4.1, predicted *apo* water positions based on pseudo-*apo* structures (based on deletion of the ligand, s. Figure 3) were compared with experimental *holo* water positions, this time predicted *holo* water positions are compared with the experimental *holo* water positions.

The respective percentages of reproduced water positions at different distance thresholds and for different B-factor subsets are given in Table 14 and are illustrated in Figure 17.

Table 14: Percentages of all experimental water positions in complexes of the used PDBbind refined subset that are correctly reproduced by the 3D RISM-based placement algorithm (using three different distance thresholds, 1.0, 1.5, and 2.0 Å). Percentages are shown for all experimental water molecules ("all") as well as for subsets of water molecules with the highest and lowest B-factors ("min X %" and "max X % B-factor"). Corresponding values for the apo structures are given in chapter 4.1 in Table 4. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure17_Table14_holo_distances_B-factors/).

distance threshold / Å	all	min 10 % B-fact.	Min 25 % B-fact.	Max 25 % B-fact.	Max 10 % B-fact.
1.0	52.3 %	75.2 %	70.8 %	34.9 %	31.3 %
1.5	77.7 %	92.0 %	89.7 %	64.9 %	60.7 %
2.0	91.6 %	97.5 %	96.8 %	85.2 %	82.8 %

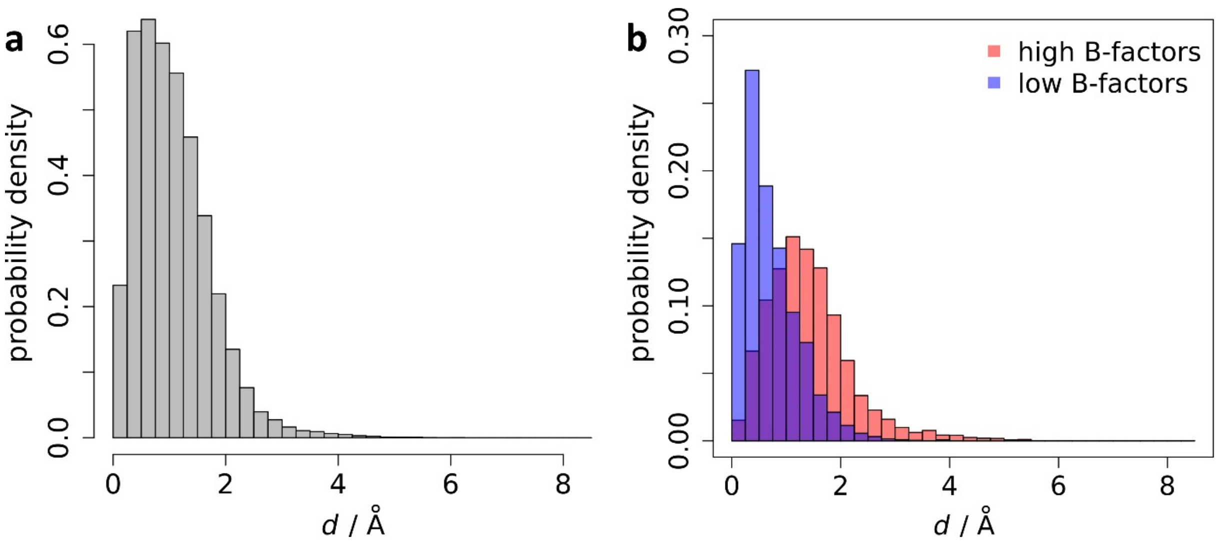

Figure 17: a) Probability densities of the distances between the experimental water positions in the complexes and the corresponding nearest calculated holo water positions as predicted by the used algorithms for all experimental water positions within the used PDBbind refined subset; b) respective probability densities for subsets of experimental water positions within the used PDBbind refined subset w.r.t. B-factor (blue: 10 % lowest B-factors, red: 10 % highest B-factors). Respective plots for the apo results are shown in 4.1 in Figure 2 *and Figure* 4*. The probability density has the inverse unit of the x-axis parameter, i.e. 1/Å. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure17_Table14_holo_distances_B-factors/).*

As expected, the agreement between predicted and experimental positions is even slightly better than for the *apo* calculations (Table 4) because the presence of the ligand is considered in the *holo* calculations): 78 % of experimental water positions are within 1.5 Å of a predicted *holo* water position (versus 74 % for the predicted apo hydration sites).The trend for better reproduction of experimental water molecules with low B-factors, too, is even slightly more pronounced for the *holo* calculations than for the *apo* calculations: 92 % of the experimental water molecules with the 10 % lowest B-factors are reproduced within 1.5 Å (versus 90 % for predicted *apo* positions).

4.1.2.2 *Holo* water thermodynamics

Upon ligand binding, several water molecules get sterically replaced by ligand atoms. Apart from this, the presence of the ligand of course also has an impact on the remaining water molecules' positions and thermodynamic properties. For instance, certain water molecules might get isolated from a former water network that was disrupted, while others might undergo favourable interactions with a polar ligand group. Analysing this altered water environment within a *holo* binding site yields important information for further ligand design, i.e., it can reveal water molecules that were destabilised upon ligand binding and could be targeted for replacement by introducing further suitable substituents.

In order to analyse to what extent the water environment within a binding site changes upon complex formation, the calculated positions and thermodynamic properties of predicted *apo* and *holo* water molecules in the used PDBbind refined subset were compared. Binding site water molecules were defined here as all water molecules with a predicted position within 3.5 Å of any ligand atom and protein atom.

When comparing *holo* with the *apo* water positions, two categories of hydration sites can be distinguished: conserved hydration sites, where a water molecule is present in both the *apo* and *holo* form (within a given threshold, here: of 1.0 or 1.5 Å), and newly introduced hydration sites (i.e. those which do not have an *apo* counterpart within 1.0 or 1.5 Å), which can correspond to *apo* water molecules which are shifted due to the presence of the ligand. Thus, all predicted *holo* water molecules are classified as either conserved or newly introduced within the presented definition. This is of course an approximation since the binding site water molecules cannot be "tracked" from *apo* to *holo*.

In Table 15, the percentages and average $\Delta_{\text{hyd}}G_{\text{P}}$ contributions are given for the conserved (i.e. those with *apo* counterparts within the respective distance threshold) and newly introduced hydration sites (i.e. those without *apo* counterparts within the respective distance threshold) using two distance thresholds, as well as the respective values for the corresponding *apo* pendants of the conserved set. For the 1.0 Å threshold, the respective histograms are shown in Figure 18.

Table 15: Average $\Delta_{hyd}G_P$ contributions of conserved and newly introduced holo binding site water molecules in the structures of the used PDBbind refined subset using a 1.0 and 1.5 Å distance threshold. In addition, also the average $\Delta_{hyd}G_P$ contributions of the apo pendants of the conserved water positions are given. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure18_Table15_Ghyd_apo_holo_conserved/).

subset	av. $\Delta_{\text{hyd}}G_{\text{P}}$ contributions / kcal·mol^{-1}
conserved (*holo*, 1.0 Å threshold)	-0.24 ± 3.44 (53 %)
conserved (*apo* pendant, 1.0 Å threshold)	-0.72 ± 4.20
newly introduced (*holo*, 1.0 Å threshold)	-0.33 ± 2.77 (47 %)
conserved (*holo*, 1.5 Å threshold)	-0.25 ± 3.31 (69 %)
conserved (*apo* pendant, 1.5 Å threshold)	-0.74 ± 4.13
newly introduced (*holo*, 1.5 Å threshold)	-0.35 ± 2.75 (31 %)

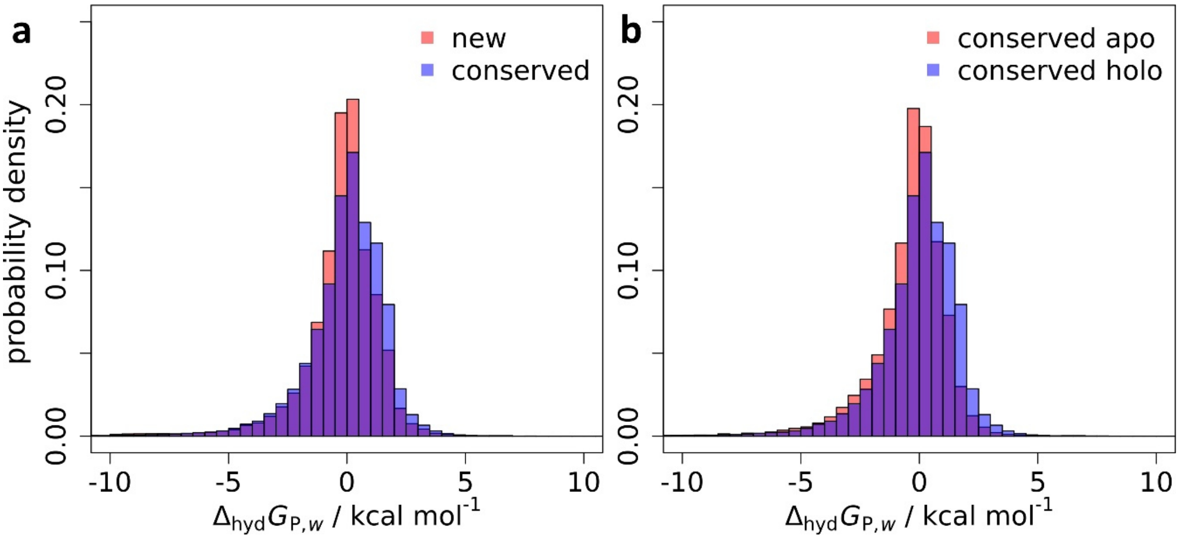

Figure 18: a) Probability densities of the average $\Delta_{hyd}G_P$ contributions of conserved (blue) and newly introduced (red) hydration sites in holo binding sites in the used PDBbind refined subset in kcal/mol; b) probability densities of the average $\Delta_{hyd}G_P$ contributions of conserved holo hydration sites (blue) and their pendant in the apo form (threshold: 1.0 Å). The probability density has the inverse unit of the x-axis parameter, i.e. kcal^{-1}·mol. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure18_Table15_Ghyd_apo_holo_conserved/).

Depending on the used threshold, 53 % to 69 % of the *holo* hydrations sites are conserved with a corresponding pendant in the *apo* form. When comparing the average $\Delta_{hyd}G_P$ contributions of the conserved and newly introduced subsets, there is a slight trend for more favourable values for the newly introduced hydration sites, implying a thermodynamically favoured shift of the *apo* water positions due to the changed environment.

Interestingly, comparison of the thermodynamic properties of the conserved *holo* water molecules with their *apo* pendants reveals that, on average, the hydration sites are slightly more "happy" in the *apo* binding sites. Reasons for this might be that their former water network was disrupted and that they can undergo less polar interactions, which also limits the number of possible orientations. For future work, it might be of interest to achieve a further dissection of solvation effects into entropic and enthalpic factors.

In a protein ligand complex, water molecules can play an important role by bridging interactions between the binding site residues and ligand groups. Therefore, the average $\Delta_{hyd}G_P$ contributions were analysed for *holo* water subsets of bridging and non-bridging water molecules. Here, a (potentially) bridging water molecule was simply defined as any predicted *holo* water molecule within 3.0 Å of both a polar ligand and a polar protein atom, neglecting preferential hydrogen bond angles. Moreover, for the non-bridging water molecules, additional subsets were defined for hydration sites which have at least one polar ligand or protein polar contact or no polar ligand or protein contact at all. The respective average $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions are given in Table 16, histograms are shown in Figure 19.

Table 16: Average $\Delta_{hyd}G_{P}$ contributions in kcal/mol of holo binding site water molecules in the structures of the used PDBbind refined subset which are considered bridging or non-bridging; the latter molecules are divided in sub groups with only polar protein contacts, only polar ligand contacts, and no polar protein or ligand contacts. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure19_Table16_holo_bridging_nonbridging/).

subset	av. $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions / kcal·mol^{-1}
bridging	-1.17 ± 5.56 (15 %)
non-bridging	-0.12 ± 2.44 (85 %)
- only protein contact	-0.35 ± 3.08 (37 %)
- only ligand contact	-0.36 ± 2.28 (12 %)
- neither	+0.19 ± 1.51 (35 %)

Within the approximation of the used definition, 15 % of the *holo* binding site water molecules are considered bridging. In average, they exhibit more favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions than the non-bridging ones. The non-bridging *holo* water molecules can be further separated into those which have either a ligand or protein polar contact or neither of them. While the hydration sites with at least one ligand or protein polar contact show comparable, slightly favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions, those with no polar contacts are in average more “unhappy”. This is in accordance with the finding in 4.1.5 that water molecules in the proximity of polar groups are more “happy” than those in an apolar environment.

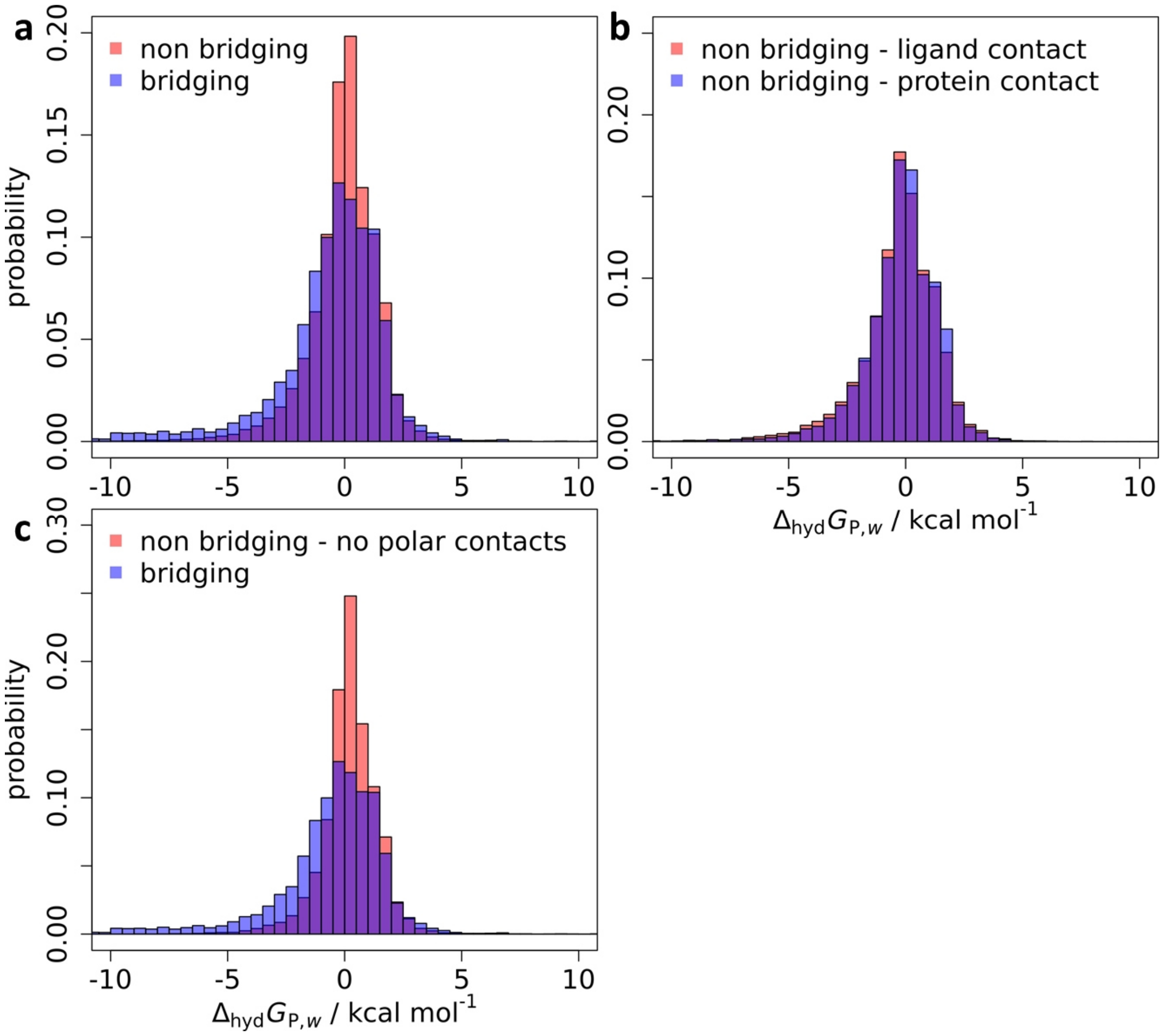

Figure 19: a) Probability densities of the average $\Delta_{hyd}G_P$ contributions of bridging (blue) and non-bridging (red) water molecules in holo binding sites in the used PDBbind refined subset; b) probability densities of the average $\Delta_{hyd}G_P$ contributions of non-bridging water molecules with only ligand polar contacts (red) and only protein polar contacts (red), c) probability densities of the average $\Delta_{hyd}G_P$ contributions of bridging water molecules (blue) and non-bridging water molecules which have neither a polar contact with a protein nor ligand atom. Bridging water molecules were defined by being within 3.0 Å of both a polar ligand and a polar protein atom, neglecting preferential hydrogen bond angles. The probability density has the inverse unit of the x-axis parameter, i.e. kcal^{-1}·mol. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure19_Table16_holo_bridging_nonbridging/).

Apart from protein and ligand atoms, also other neighbouring water molecules coin the chemical environment of a given hydration site. Therefore, the thermodynamic properties of water molecules with varying numbers of neighbouring water molecules were investigated. The respective average $\Delta_{hyd}G_P$ contributions are presented in Table 17, and a representative histogram is shown in Figure 20.

Table 17: Average $\Delta_{hyd}G_P$ contributions of holo binding site water molecules in the structures of the used PDBbind refined subset which have 1) no neighbouring water molecules, 2) no neighbouring water molecules and no polar ligand or protein contacts (subset of 1), 3) one neighbouring water molecule, and 4) two or more neighbouring water molecules, using a threshold of 3.0 Å. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure20_Table17_holo_near_waters/).

subset	av. $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions / kcal·mol^{-1}
no neighbouring water molecules	+0.17 ± 3.37 (30 %)
- no neighbouring water molecules + no polar contact	+0.59 ± 1.16 (10 %)
1 neighbouring water molecule	-0.27 ± 2.65 (37 %)
2 or more neighbouring water molecules	-0.71 ± 3.37 (33 %)

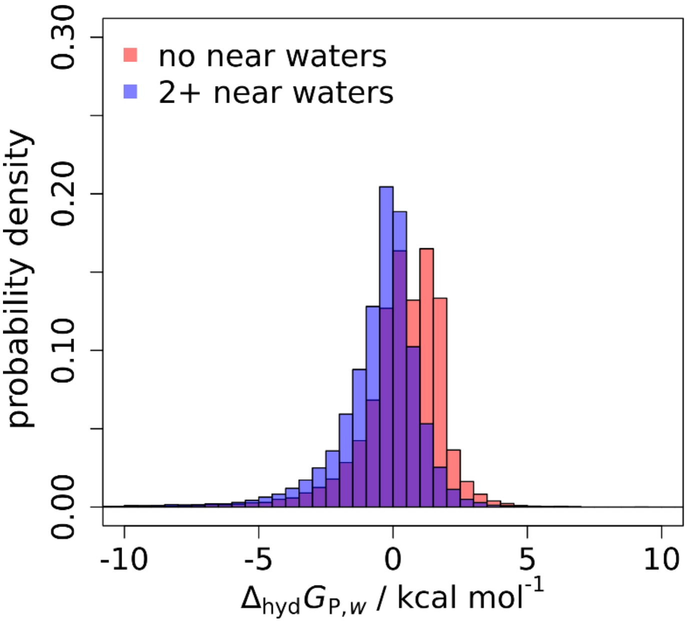

Figure 20: Probability densities of the average $\Delta_{hyd}G_P$ contributions of holo binding site water molecules with no neighbouring water molecules (red) and with two or more neighbouring water molecules (blue), using a threshold of 3.0 Å. The probability density has the inverse unit of the x-axis parameter, i.e. kcal^{-1}·mol. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ data/ PDBbind_refined_set/ data/ Figure20_Table17_holo_near_waters/).

In general, water molecules with two or more neighbouring water molecules, which make up one third of all binding site water molecules, show the most favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions. For hydration sites with only one neighbouring water molecule, the average $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contribution is still slightly favourable, while those with no other water molecules in their proximity in average exhibit slightly unfavourable

$\Delta_{\text{hyd}}G_{\text{P}}$ contributions. The most pronounced trend can be observed for a subset of water molecules which have neither neighbouring water molecules nor polar ligand or protein contacts. The identification of such isolated, high-energy hydration sites is of relevance for drug design purposes since their replacement by a slightly modified ligand is likely highly beneficial for binding affinity, as will also be shown in chapters 4.1.3, 4.3.1, and 4.3.3. In future work, it could be of interest to further dissect the solvation effects into enthalpic and entropic contributions since entropy is likely a determining factor for rather isolated binding site water molecules. Steps into this direction have already been made using GIST;[231] however, in this case, the entropy is described by solute-water correlations only (ΔS_{uv}), neglecting the water-water-correlations (ΔS_{vv}). Hence, a 3D RISM-based approach including both terms might be desirable in the future to get a deeper insight why a specific water molecule is "happy" or "unhappy".

4.1.2.3 Influence on ligand affinity

In chapter 4.1, the correlation of *apo* water thermodynamics and ligand affinity was investigated. The respective studies revealed that, in general, the druggable binding sites (which are represented within the data set) contain more high energy hydration sites than undruggable ones, and that high affinity ligands tend to replace more "unhappy" water molecules. Furthermore, specific replacement propensities of certain ligand functional groups were observed, for instance that hydroxyl groups preferentially replace "happy" water molecules, while halogen atoms and aromatic carbon atoms replace rather "unhappy" water molecules.

In this chapter, it will be investigated if any trends can be observed w.r.t. *holo* water properties and ligand affinity. For instance, one might assume that an increase in the number of high energy water molecules is unfavourable for binding - however, this might of course be compensated by respective interactions. Likewise, the presence of a lot of bridging waters could be beneficial in terms of polar interactions but might go along with an isolation or oriental restriction of the respective water molecules.

Besides, the used PDBbind refined subset contains a large variety of complexes of different protein classes and ligands of varying size and chemical space, even for the same protein, which makes an analysis difficult. Therefore, some important aspects have to be considered when trying to draw conclusions: As indicated, in 4.1.1.7 it was shown that binding sites which are considered druggable exhibit a higher fraction of water molecules with unfavourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions, and that the ratio of "unhappy" *apo* water molecules correlates with higher ligand affinity. Therefore, even in the presence of the ligand, the more druggable binding sites with the ligands of higher affinity will have an increased number of "unhappy" water molecules, so that parameters like the average $\Delta_{\text{hyd}}G_{\text{P}}$ contribution of all binding site water molecules, or the ratio of "unhappy" binding site water molecules among all hydration sites are not well-suited to capture trends w.r.t. ligand affinity. Rather, a measure is needed that captures the changes in binding site water thermodynamics due to the presence of the ligand.

Here, the parameter $\Delta N_{\mathrm{unhappy},i}$ is introduced which measures the change in the number of "unhappy" water molecules from the *apo* to the *holo* form of a given protein *i* according to:

$$\Delta N_{\mathrm{unhappy},i} = N_{\mathrm{unhappy},holo,i} - N_{\mathrm{unhappy},apo,i} \tag{71}$$

Here, $N_{\mathrm{unhappy},holo,i}$ and $N_{\mathrm{unhappy},apo,i}$ denote the total number of binding site water molecules whose $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contribution is >= 0.0 kcal/mol ("unhappy" binding site water molecules) in the *apo* and *holo* form of a given protein *i*. Due to the used definition of binding site water molecules (within 3.5 Å of any ligand and protein atom), the total number of binding site water molecules increases with ligand size, which itself correlates with ligand affinity. To obtain a relative measure, $\Delta N_{\mathrm{unhappy},i}$ can be divided by the total number of *holo* binding site water molecules in a given protein *i*, $N_{holo,i}$, yielding the ratio $\Delta x_{\mathrm{unhappy},i}$ according to:

$$\Delta x_{\mathrm{diff,unhappy},i} = \frac{\Delta N_{\mathrm{unhappy},i}}{N_{holo,i}} \tag{72}$$

In contrast to the analysis in 4.1.1, the measures $\Delta N_{\mathrm{unhappy},i}$ and $\Delta x_{\mathrm{unhappy},i}$ do not only capture the steric replacement of "unhappy" *apo* binding site water molecules, but also the possible introduction of new high energy hydration sites due to the presence of the ligand. The latter could potentially compensate or overweigh the beneficial effects of the *apo* waters' replacement.

In Table 18, the *R*- and *p*-values for the correlation of $\Delta N_{\mathrm{unhappy}}$ and $\Delta x_{\mathrm{unhappy}}$ with ligand affinity on the whole used PDBbind refined subset are given (scatterplots in Appendix, 7.5) together with exemplary average values of $\Delta N_{\mathrm{unhappy}}$ and $\Delta x_{\mathrm{unhappy}}$ for subsets containing only complexes with the 10 % and 25 % most and least affine ligands. In addition, for comparison, the same values are given for x_{unhappy}, simply defined as the ratio of "unhappy" water molecules among all *holo* binding site water molecules using the same thresholds. Respective histograms for the 10 % most and least affine ligands are shown in Figure 21.

Table 18: R- and p-values of the Pearson correlation of $x_{unhappy}$, the ratio of "unhappy" water molecules, $\Delta N_{unhappy}$, and $\Delta x_{unhappy}$ as defined in Eq. (71) and (72) with ligand affinity for the whole used PDBbind refined set. In addition, average values of the parameters are given for subsets of complexes with ligands with the X % highest and lowest affinity. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure21_Table18_diff_unhappy/).

parameter	R	p	max 10 %	max 25 %	min 25 %	min 10 %
			p$K_{\text{aff.}}$ > 9.0	p$K_{\text{aff.}}$ > 7.9	p$K_{\text{aff.}}$ < 5.0	p$K_{\text{aff.}}$ < 3.8
x_{unhappy}	0.08	< 0.01	0.59 ± 0.21	0.56 ± 0.22	0.51 ± 0.27	0.50 ± 0.26
$\Delta N_{\text{unhappy}}$	-0.20	< 0.01	-3.57 ± 5.31	-2.97 ± 5.24	-0.30 ± 3.88	-0.13 ± 3.47
$\Delta x_{\text{unhappy}}$	-0.14	< 0.01	-0.24 ± 0.42	-0.19 ± 0.41	-0.05 ± 0.33	-0.04 ± 0.32

The results show that x_{unhappy}, the ratio of "unhappy" water molecules within a given binding site, is in average higher for complexes with ligands of high affinity. As discussed above, this is an indirect effect because druggable binding sites contain more "unhappy" water molecules in general. For $\Delta N_{\text{unhappy}}$ and $\Delta x_{\text{unhappy}}$, however, a different trend is observed: For both parameters, there is a slight, yet significant anti-correlation with affinity which can also be seen in the respective histograms for the subsets with the most and least affine ligands (Figure 21). This implies that, in the high affinity complexes, a larger net reduction of the number of "unhappy" water molecules is achieved from the *apo* to the *holo* form. This anti-correlation is more pronounced for the absolute count $\Delta N_{\text{unhappy}}$ than for the relative ratio $\Delta x_{\text{unhappy}}$ (R = -0.20 vs. -0.14), which might result from fact that larger ligands, which in average are more affine, indeed replace more "unhappy" water molecules. However, overall, the correlation of all studied parameters with ligand affinity is rather weak since other factors, like intermolecular interactions or internal ligand energy easily outweigh the effect of the reduction of the number of "unhappy" water molecules.

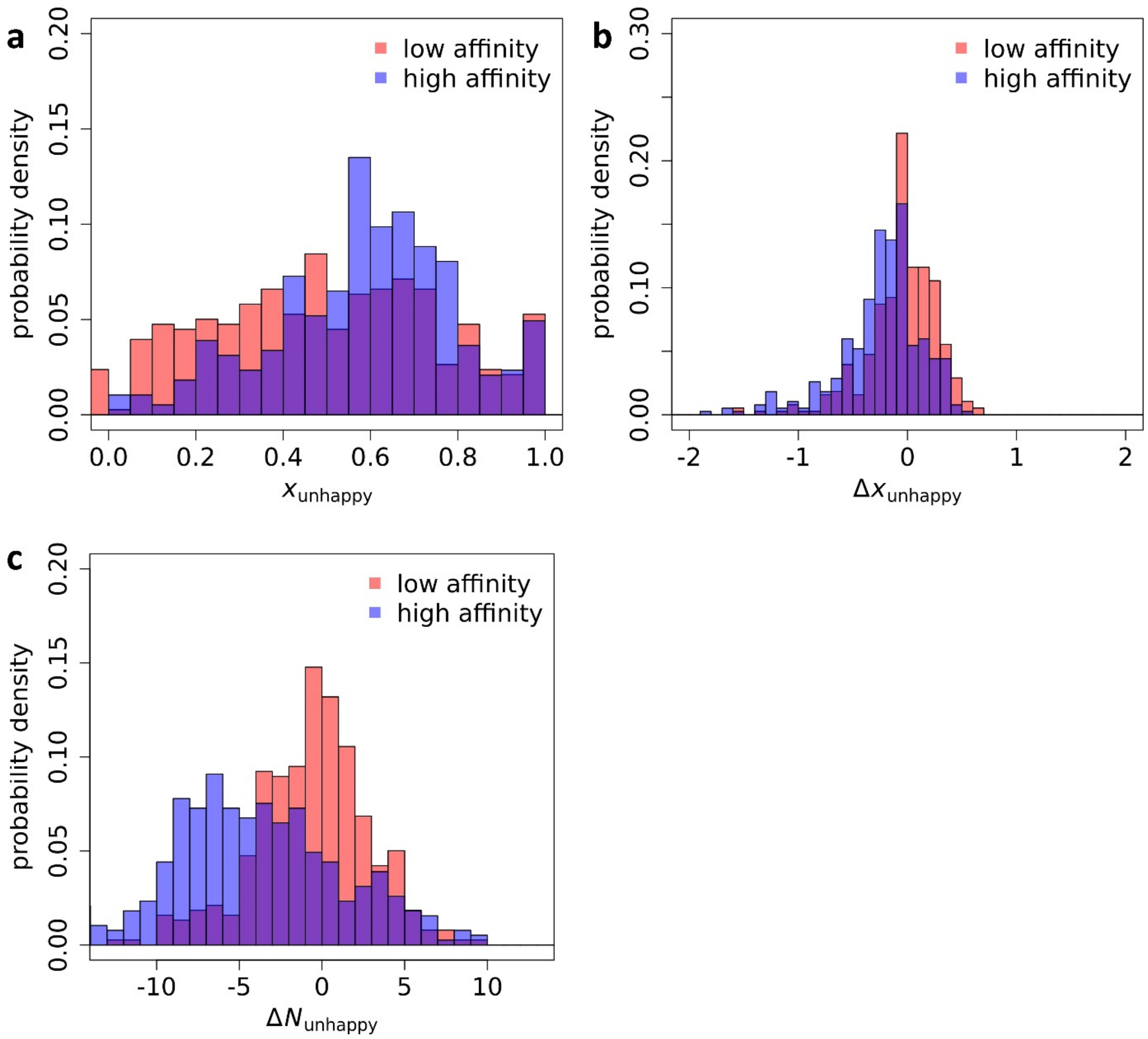

Figure 21: Probability densities of a) $x_{unhappy}$, the ratio of "unhappy" water molecules among all holo binding site water molecules, b) $\Delta x_{unhappy}$, and c) $\Delta N_{unhappy}$ as defined in Eq. (71) and (72) for subsets containing only complexes with the 10 % most (blue) and 10 % least (red) affine ligands within the used PDBbind refined subset. The probability density has the inverse unit of the x-axis parameter. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Figure21_Table18_diff_unhappy/).

So far, all presented analyses cover the whole PDBbind refined set, or affinity subsets of it, to derive universally valid trends w.r.t. water thermodynamics. However, the large number of complexes in the data set potentially also allows for an investigation of trends within specific protein classes. Therefore, selected proteins were studied which are represented within the used PDBbind refined subset by at least 25 complex structures and whose ligands cover a broad affinity range (>= 5 pK_i/pK_d units). The *R*- and *p*-values for the Pearson correlation of x_{unhappy}, $\Delta N_{\text{unhappy}}$, and $\Delta x_{\text{unhappy}}$ with ligand affinity for the respective protein subsets are given in Table 19. The corresponding scatterplots can be found in the Appendix (7.5.2).

Table 19: Average number of heavy ligand atoms (N_{HA}), as well as R- and p-values for the Pearson correlation of $x_{unhappy}$, the ratio of "unhappy" water molecules among all holo binding site water molecules, $\Delta N_{unhappy}$,, and $\Delta x_{unhappy}$ as defined in Eq. (71) and (72) with ligand affinity for selected proteins within the used PDBbind refined subset (and ** denote significant and highly significant correlations). Corresponding scatterplots can be found in the Appendix (7.5.2). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ Table19_protein_groups/).*

		$x_{unhappy}$		$\Delta N_{unhappy}$		$\Delta x_{unhappy}$	
	N_{HA}	*R*	*p*	*R*	*p*	*R*	*p*
BACE1	31 ± 13	-0.55**	< 0.01	-0.27	0.10	-0.25	0.13
BRD4	27 ± 5	-0.27	0.06	-0.22	0.11	-0.16	0.27
CA2	18 ± 6	0.21**	< 0.01	0.08	0.20	-0.01	0.83
caseinkinaseII	20 ± 6	-0.26	0.10	0.48**	< 0.01	0.54**	< 0.01
fXa	32 ± 3	0.27	0.09	0.19	0.24	0.25	0.13
HIV1PR	44 ± 7	0.32**	< 0.01	0.19**	< 0.01	0.11	0.09
HSP90	24 ± 6	-0.20	0.06	-0.22*	0.03	-0.19	0.08
MMP12	29 ± 12	0.42*	0.03	0.38	0.05	0.44*	0.02
NA	20 ± 2	0.44*	0.02	-0.33	0.09	-0.32	0.11
thermolysin	27 ± 6	-0.32	0.09	-0.56**	< 0.01	-0.60**	< 0.01
thrombin	27 ± 5	-0.30	0.11	-0.30	0.11	-0.22	0.25
trypsin	22 ± 10	-0.06	0.64	-0.24	0.07	0.00	1.00

Unfortunately, almost no significant correlations can be observed for the protein subsets. One reason for this likely is the small size of each subset. However, the main reason for the lack of significance probably is that the results are biased if complexes of the same protein contain ligands of strongly differing size or ligands that occupy different parts of a binding site (due to the definition of binding site water molecules via a distance criterion from the ligand). Besides, the provided subsets usually do not contain ligand series of highly related structures but a variety of ligands with different scaffolds.

Yet, certain differences can be observed both between the different protein subsets and in comparison with the statistics on the whole data set: For the proteins BACE1, BRD4, HSP90, thermolysin, and thrombin, a higher ligand affinity does not only correlate with more negative values of $\Delta N_{\mathrm{unhappy}}$ and $\Delta x_{\mathrm{unhappy}}$, i.e. a higher net replacement of "unhappy" water molecules, but also with a lower ratio of "unhappy" water molecules in the *holo* binding sites. This is opposed to the overall trend of a higher "unhappy" water ratio for high affinity ligands. Yet, it is intuitive because the binding site heterogeneity that leads to a non-uniform distribution of ligand affinities between the different proteins is neglected when regarding only one protein. Thus, when comparing ligands that bind in roughly the same area, it is reasonable that the presence of less "unhappy" water molecules in the proximity of the ligands can correlate with higher affinity.

Indeed, BACE1 is known as a prominent example for which water thermodynamics play an important role in ligand SAR. A study by Brodney *et al*. revealed a strong correlation between ligand affinity and the free energy liberation of binding site water molecules as calculated by WaterMap for a series of spiropiperidine ligands.[281] In the used PDBbind refined subset, BACE1 is represented by 38 complex structures whose ligands however vary in size and are not structurally closely related, so that only rough trends can be observed, with the correlation of x_{unhappy} with binding affinity being the only one that is statistically significant. Nevertheless, the observed tendencies are in line with the trends observed in literature, highlighting the relevance of binding site water molecules for drug design on BACE1.

Thermolysin is another interesting target in this context. Studies revealed that replacement of waters from imperfectly hydrated pockets is favourable for binding.[282,283] This is captured by the strong and highly significant correlation of $\Delta N_{\text{unhappy}}$ and $\Delta x_{\text{unhappy}}$ with ligand affinity, which is more pronounced than for any of the other protein subsets. Visual inspection of the respective complex structures revealed that most ligands in the thermolysin subset share a phosphonamidate backbone and occupy similar areas of the binding sites (Figure 22a), which makes this subset ideally suited for a comparative analysis.

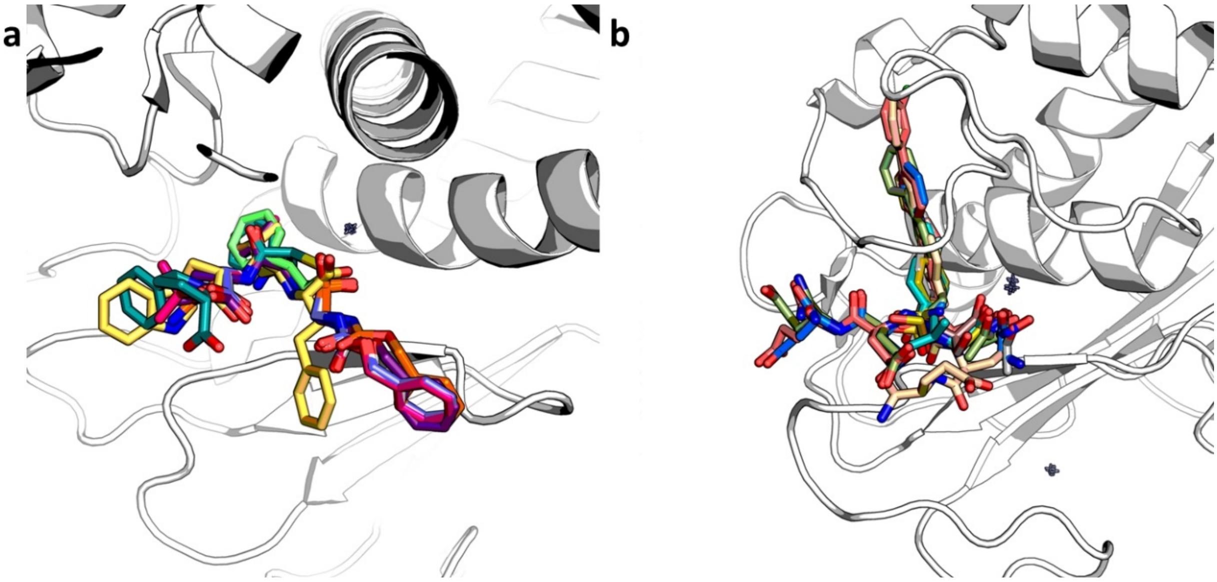

Figure 22: Overlay of representative ligands in the used protein subsets for a) thermolysin (5m9w,5ma7,[284] 5n2t,[285] 1zdp,[286] 1qf2,[287] 5jss,[206] 1tmn,[288] 5tmn;[289] cartoon: 5m9w[284]), and b) MMP12 (3lka,[290] 3lir,[291] 3f18,[292] 3ljg,[291] 1rmz,[293] 3f16,[292] 3ts4,[294] 5d3c;[295] cartoon: 3lka[290]).

An illustrative example of a Matched Molecular Pair (MMP) within the data set is shown in Figure 23. The ligand in complex structure 5tmn differs from the one in 5m9w only in the addition of an isopropyl substituent which occupies a pocket untargeted by the ligand in 5mw9. As the interpolated *apo* $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions reveal, it replaces highly unstable hydration sites, leading to a massive gain in affinity (5m9w: pK_{aff} = 2.24, 5tmn: pK_{aff} = 8.04).

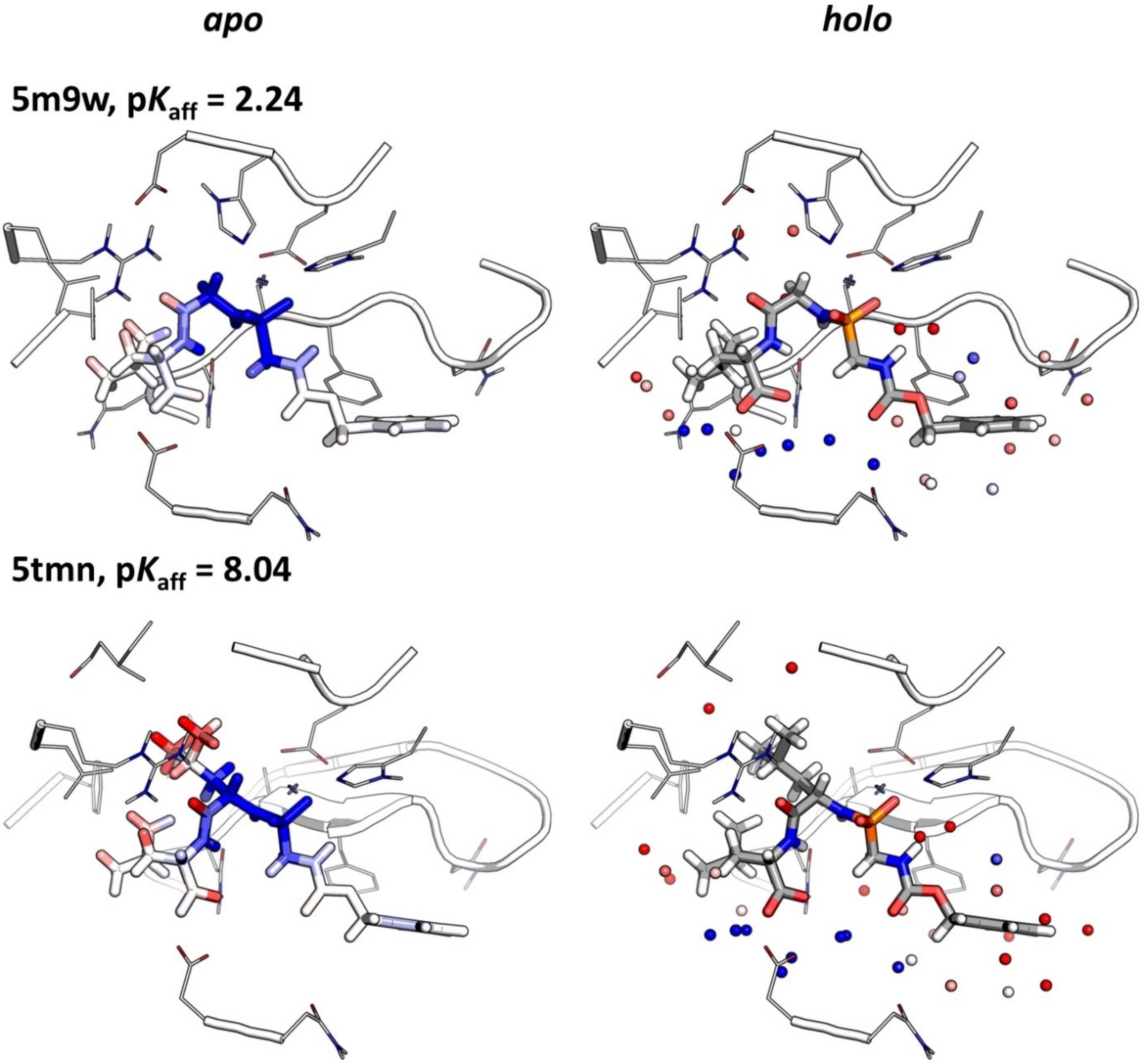
apo
holo
5m9w, pK_{aff} = 2.24
5tmn, pK_{aff} = 8.04

Figure 23: Thermolysin MMP in complex structures 5m9w[284] *and 5tmn:*[289] *Left: ligand, coloured by the interpolated $\Delta_{hyd}G_P$ contributions of replaced apo water molecules (from blue to red from -2.0 to +2.0 in units of kcal/mol); right: ligand with predicted holo water molecules (within 3.5 Å of any ligand atom), coloured by their calculated $\Delta_{hyd}G_P$ contributions (same colour code). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ structures/) in the respective pdb folders.*

The clear trend for the thermolysin subset emphasises that the analysis of water thermodynamics is well suited to help explain SAR trends for a set of structurally related ligands of a given protein. However, as stated before, these water thermodynamics-related trends are easily outweighed by other effects. Therefore, no or even an opposed tendency can be seen for the proteins CA2, fXa, HIV1PR, MMP12, and NA, since the protein-specific ligand subsets used here are rather small and diverse. For MMP12, for instance, ligands vary strongly in size and occupy different parts of the pocket (Figure 22b). For a meaningful analysis for specific protein classes, larger data sets with closely related ligands would be needed. Such data sets – especially with crystal structures of all related ligands – are most likely to be generated during drug development programmes in pharmaceutical industry but unfortunately are often not publicly available.

Yet, the strong correlations for BACE1 and thermolysin, two prominent examples with known relevance of water replacement for ligand design, underline that 3D RISM based analysis of *apo* and *holo* structures can yield strategies for ligand improvement. Therefore, in the next chapter, both the *apo* and *holo* analyses will be exploited for explaining SAR trends for a set of MMPs within the used PDBbind refined subset.

4.1.3 Water thermodynamics and SAR trends – MMP case studies

In Chapter 4.1 and 4.2, the characteristics of water molecules in *apo* and *holo* protein binding sites were studied using a large data set, and valuable conclusions were drawn w.r.t. ligand design. While the afore described analysis was focused on general trends and statistics, the following chapter will present specific case studies of MMPs within the used PDBbind refined set for which water thermodynamics can be employed to explain respective SAR trends. Special emphasis will be put on the replacement rules derived in 4.1.10. Here, it was shown that rather hydrophobic groups practically exclusively replace "unhappy" water molecules – since the penalty of replacing a more "happy" water molecule cannot be sufficiently compensated by favourable, strong interactions with the binding site residues. "Happy" waters, on the other hand, tend to get replaced by polar groups, likely because their microenvironment is well suited to accommodate such ligand moieties. A particularly relevant finding was found for hydroxyl groups – the analysis in 4.1.10 revealed that, in high affinity ligands, this functional group scarcely replaces any "unhappy" water molecules, thus suggesting that primarily "happy" water molecules should be targeted by a hydroxyl moiety. The analysis that is presented in this chapter provides concrete examples to illustrate and complement the general trends derived by the described large-scale analysis.

In the following examples, MMPs are discussed by investigating the structural complex data with both the *apo* and *holo* water positions and thermodynamics. In the respective illustrations, the less affine MMP complex is always shown on the left side of the panel, the more affine one on the right. The *apo* water positions and thermodynamics are shown in the top row, the respective *holo* data in the bottom row. For the *apo* form, the ligands are additionally coloured according to the interpolated *apo* $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions. In the top right panel, areas of special interest (for instance, where MMPs differ or where a water molecule with an especially favourable or unfavourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contribution is located) are highlighted with a green circle.

When looking at the examples, one has to keep in mind that the presented analysis is limited to the solvation part and completely neglects direct interactions or the respective ligand conformation. Hence, the respective analysis should not be over-interpreted but should rather be seen as one part of the explanation for the affinity differences in the presented ligands.

Figure 24 shows an MMP of α-mannosidase II. Both ligands do only differ in an additional methyl group at the nitrogen atom in the 5-ring, yet the change in affinity is remarkable (3ddf:[296] pK_{aff} = 4.66, 3ddg:[296] pK_{aff} = 6.00). Water thermodynamics can help to explain this notable gain in affinity: While the *apo* water molecules in the proximity of the catalytic Zn ion show highly favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$

contributions, the region which accommodates the additional methyl group in 3ddg contains unstable hydration sites. The methyl group is able to efficiently replace these “unhappy” water molecules while the hydrogen atom in 3ddf is too small. This is also represented in the predicted water thermodynamics of the *holo* form: In 3ddf, a high energy water molecule is still located next to the respective hydrogen of the pyrrolidine ring at the same position as in the *apo* form. In 3ddf, on the other hand, the water molecules in the respective area are eliminated due to the presence of the larger methyl group.

Intriguingly, there is another MMP of α-mannosidase II ligands in the used PDBbind refined set which is related to a similar substitution in the very same region of the binding site. As can be seen in Figure 25, the more affine ligand in 3dx2[297] bears an additional methylsulfinyl moiety corresponding to the position of the methyl group in 3ddg. The methylsulfinyl group replaces several high energy water molecules, while the unsubstituted ring system is not large enough, so that “unhappy” water molecules remain in the respective binding site part. The achieved gain in affinity for this MMP is even more pronounced (3dx1:[297] pK_{aff} = 3.58, 3dx2: pK_{aff} = 6.82) than for 3ddf/3ddg, which impressively highlights that small structural changes in a ligand can have a tremendous impact on binding affinity when optimally adjusted to the binding site environment.

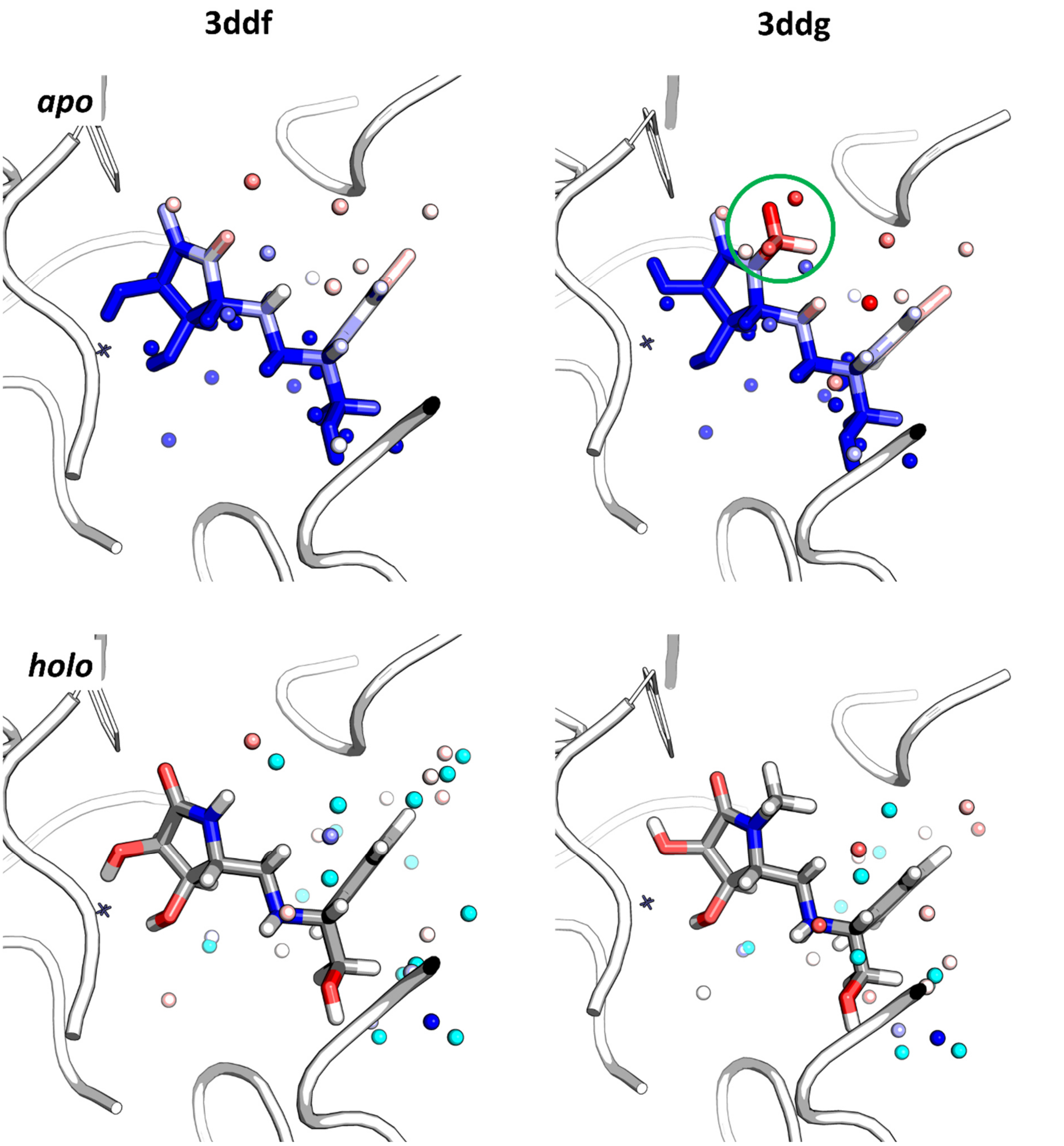

Figure 24: Complex structures 3ddf (pK$_{aff}$ = 4.66) and 3ddg (pK$_{aff}$ = 6.00) with apo and holo water positions as predicted by 3D RISM-based algorithms, coloured by their $\Delta_{hyd}G_P$ contributions, and experimental holo water positions (cyan). For the apo form, the ligand is coloured by the respective interpolated $\Delta_{hyd}G_P$ contributions of replaced apo water molecules (colouring from blue to red from -2.0 to +2.0 in units of kcal/mol). Water positions within 3.5 Å of the ligand are shown. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ structures/) in the respective pdb folders.

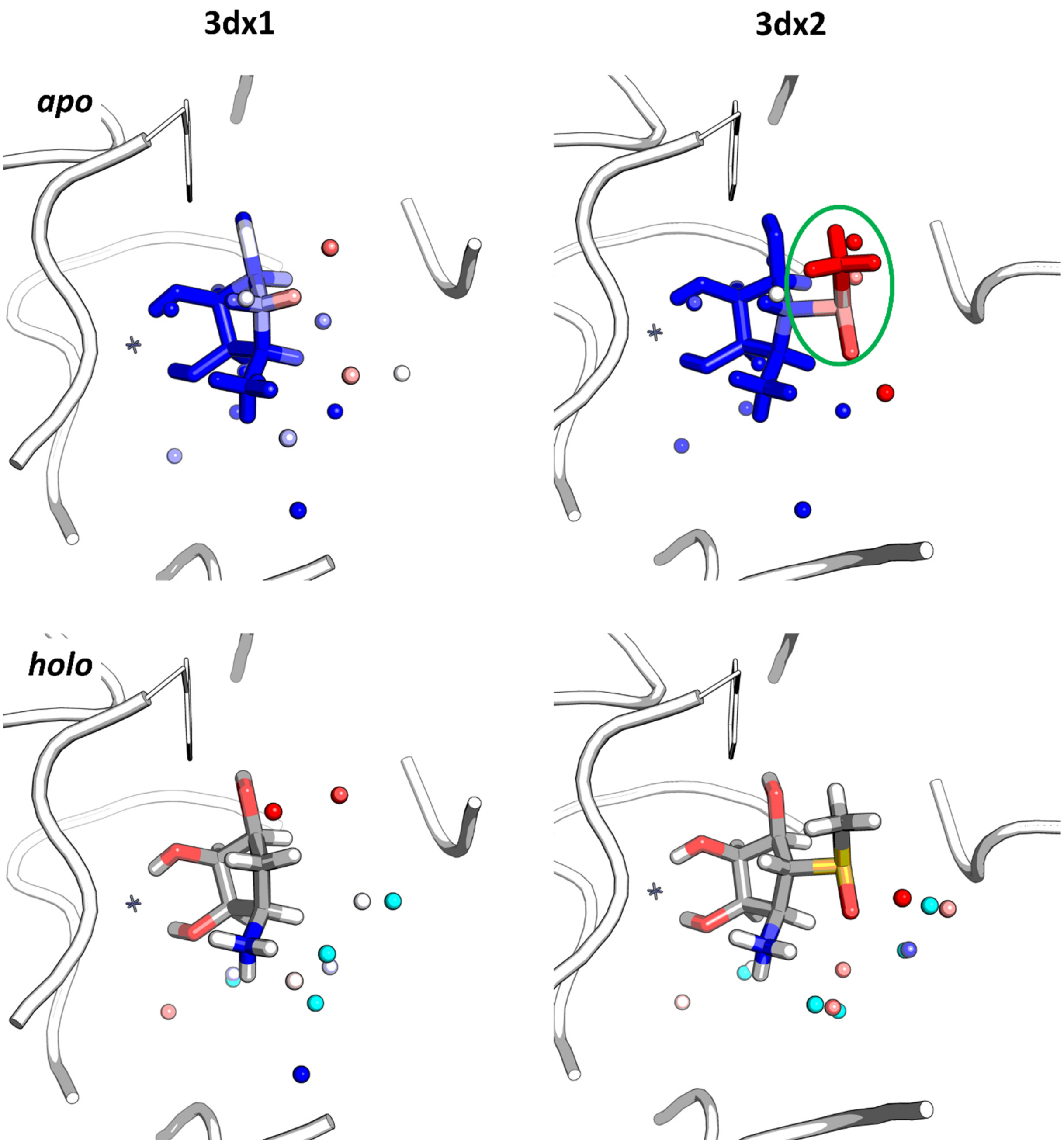

Figure 25: Complex structures 3dx1 (pK$_{aff}$ = 3.58) and 3dx2 (pK$_{aff}$ = 6.82) with apo and holo water positions as predicted by 3D RISM-based algorithms, coloured by their $\Delta_{hyd}G_P$ contributions, and experimental holo water positions (cyan). For the apo form, the ligand is coloured by the respective interpolated $\Delta_{hyd}G_P$ contributions of replaced apo water molecules (colouring from blue to red from -2.0 to +2.0 in units of kcal/mol). Water positions within 3.5 Å of the ligand are shown. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ structures/) in the respective pdb folders.

In Figure 26, an MMP of adenosine deaminase is shown which, too, illustrates the relevance of the replacement of unstable hydration sites. In the more affine ligand, a phenyl substituent is replaced by a larger naphthalene moiety (1ndw:[298] pK_{aff} = 5.23; 1ndy:[298] pK_{aff} = 6.17). Comparison of the respective *apo* water thermodynamics reveals that the region where the phenyl and naphthalene moieties bind are occupied by several high energy water molecules. The larger naphthalene moiety can more efficiently replace them, as can be seen nicely in the interpolated *apo* $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions. For the *holo* complexes, no significant differences can be seen when comparing both complexes; due to the presence of the hydrophobic ligand groups, some “unhappy” water molecules are present at the ridge of the binding site. Thus, in this MMP, the replacement of a higher number of unfavourable *apo* hydration sites likely is, among others (like i.e. the possibility to undergo stronger van der Waals interactions), one factor that leads to the increased affinity of the 1ndy ligand.

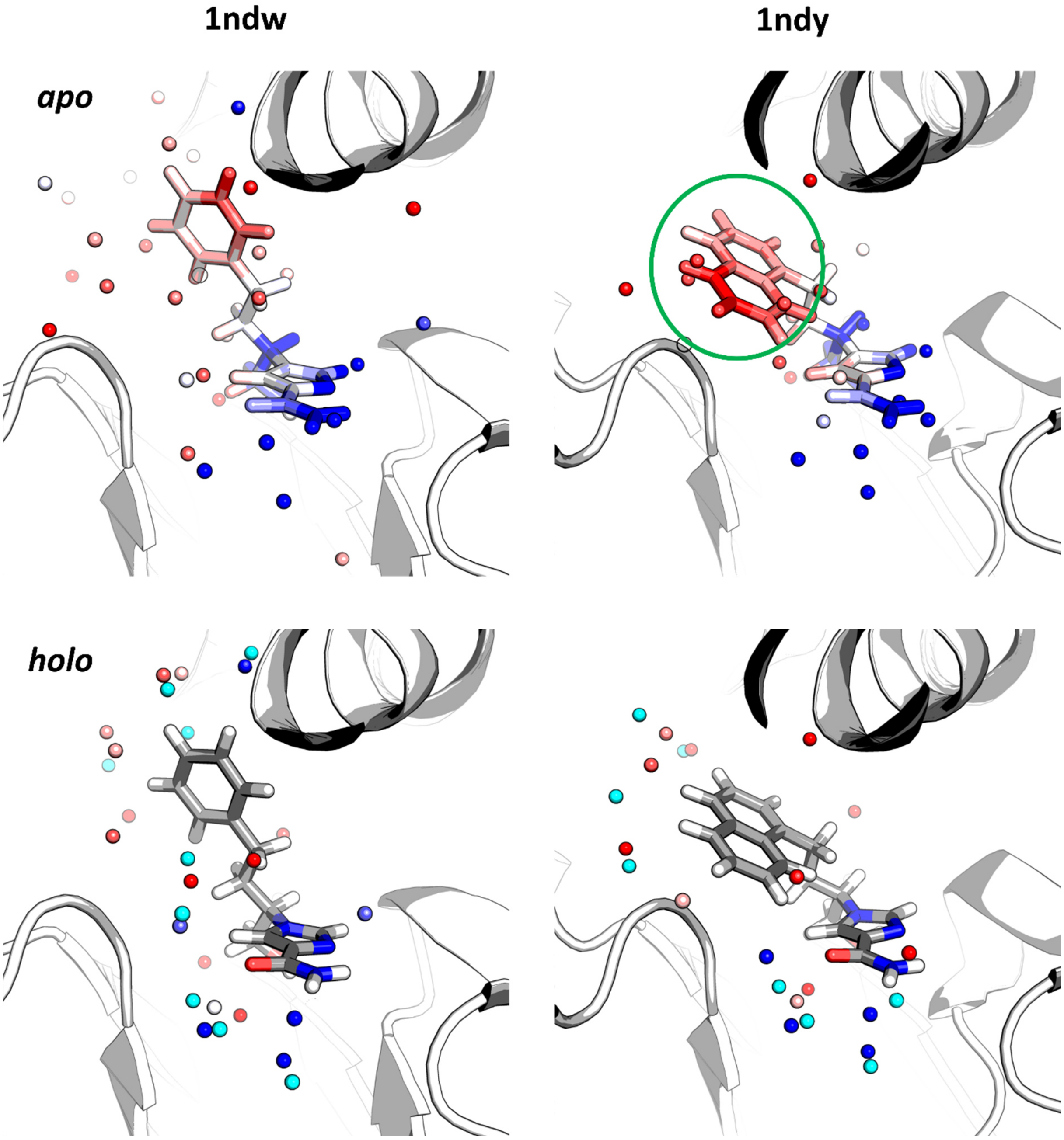

Figure 26: Complex structures 1ndw (pK$_{aff}$ = 5.23) and 1ndy (pK$_{aff}$ = 6.17) with apo and holo water positions as predicted by 3D RISM-based algorithms, coloured by their $\Delta_{hyd}G_P$ contributions, and experimental holo water positions (cyan). For the apo form, the ligand is coloured by the respective interpolated $\Delta_{hyd}G_P$ contributions of replaced apo water molecules (colouring from blue to red from -2.0 to +2.0 in units of kcal/mol). For the 1ndw apo form, water molecules within 6 Å of the ligand are shown for better comparison, otherwise within 3.5 Å. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ structures/) in the respective pdb folders.

In Figure 27, an interesting MMP of fXa is shown for which the same “unhappy” water molecules are replaced by groups with different properties: The more affine ligand (2xbv:[87] pK_{aff} = 8.43) contains a difluoroethyl substituent at the pyrrolidine ring, while the less affine ligand (2xbx:[87] pK_{aff} = 7.82) bears a methylsulfonyl substituent at the respective position. Hence, both groups replace the unstable hydration sites located in this area. Likely, the respective environment is better suited to accommodate the less polar difluoroethyl group, which is in line with the finding that, in the more affine ligands, halogen atoms tend to replace especially “unhappy” water molecules while sulfones are enriched in the proximity of “happy” water molecules. However, in this case the affinity difference is rather small.

Another fXa MMP worth discussing is presented in Figure 28. As already outlined by the Kast working group in earlier work,[192] the S1 pocket of fXa contains a highly unstable hydration site whose replacement is beneficial for affinity. This can also be seen for the complexes in Figure 27, where the respective hydration site is replaced by a chlorine residue at the pyridine ring. The two complexes shown in Figure 28 (2bq7: pK_{aff} = 7.05, 2boh: pK_{aff} = 8.52) present an example, known in literature,[299] how binding affinity is decreased if the respective ligand group is too small for completely desolvating the S1 pocket: While the more affine ligand in 2boh contains an isoxazole and a thiophene ring with a chlorine substituent, the less affine ligand in 2bq7 only bears a methoxyphenyl moiety in this region. Consequently, the respective branch is shorter, so that the methoxyphenyl group cannot fully occupy the S1 pocket. As can be seen for the *holo* forms, this leads to the presence of an isolated, unstable hydration site in the proximity of the methoxy group in the S1 pocket for 2bq7, while no retained *holo* water position is predicted for 2boh. This is in accordance with the experimental structural data: the respective crystallographically determined water molecule observed near the methoxy group in 2bq7 (shown in cyan) is in good agreement with the predicted water position. Like the α-mannosidase II ligands, this example highlights the potential large impact of a single, unstable hydration site for binding affinity.

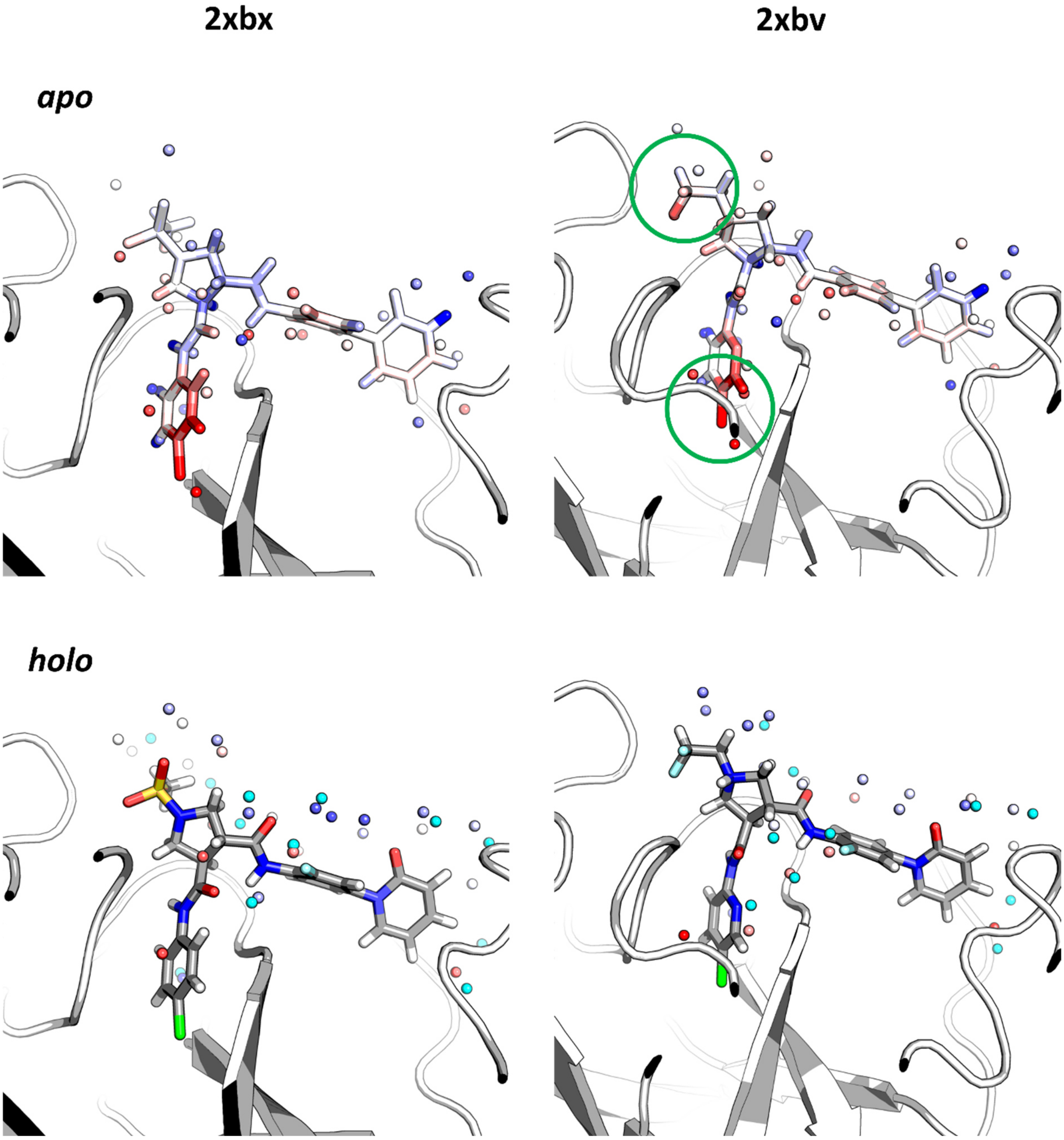

Figure 27: Complex structures 2xbx (pK_{aff} = 7.82) and 2xbv (pK_{aff} = 8.43) with apo and holo water positions as predicted by 3D RISM-based algorithms, coloured by their $\Delta_{hyd}G_P$ contributions, and experimental holo water positions (cyan). For the apo form, the ligand is coloured by the respective interpolated $\Delta_{hyd}G_P$ contributions of replaced apo water molecules (colouring from blue to red from -2.0 to +2.0 in units of kcal/mol). Water positions are shown within 3.5 Å of the ligand. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ structures/) in the respective pdb folders.

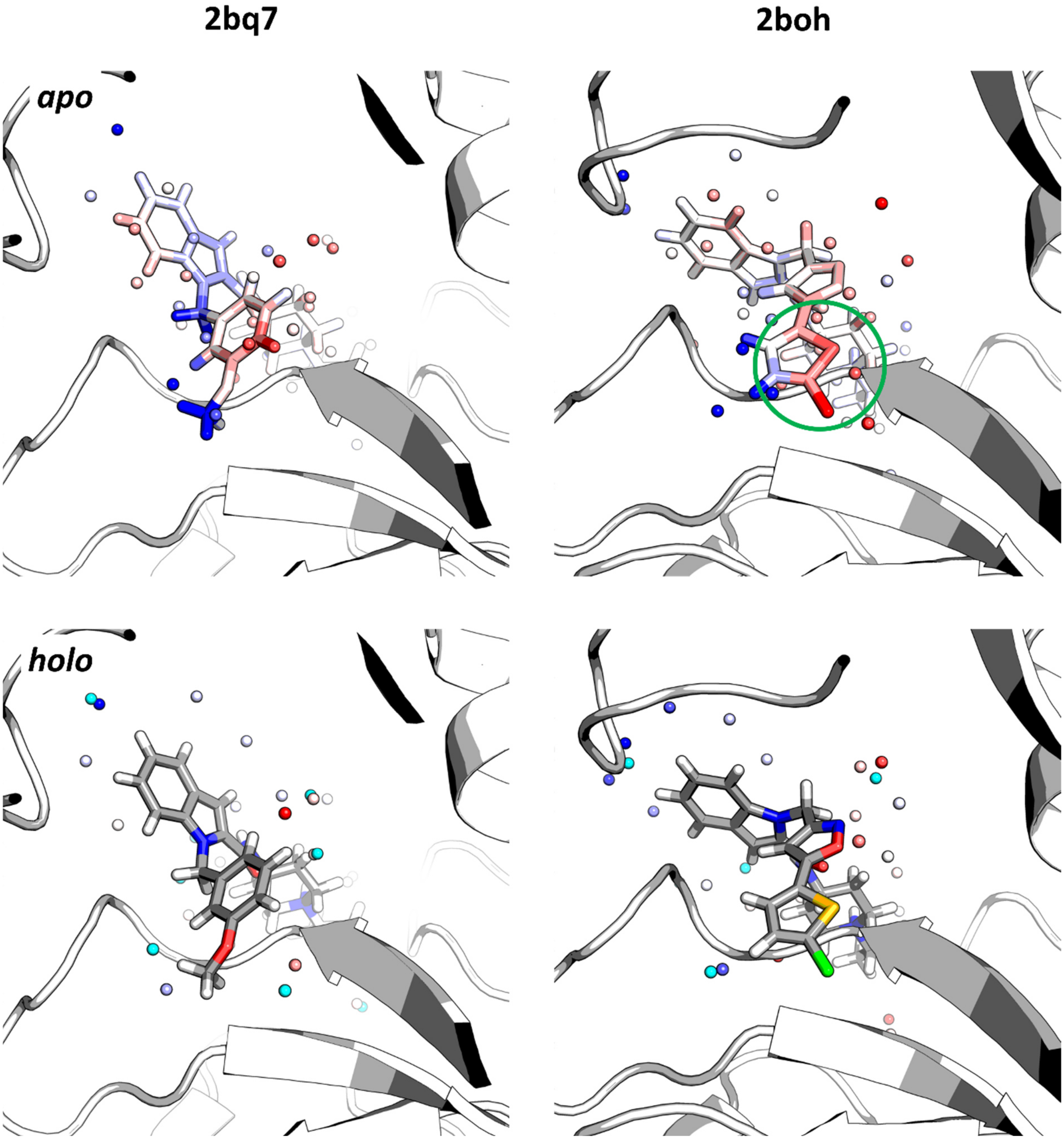

Figure 28: Complex structures 2bq7[300] (pK_{aff} = 7.05) and 2boh[300] (pK_{aff} = 8.52) with apo and holo water positions as predicted by 3D RISM-based algorithms, coloured by their $\Delta_{hyd}G_P$ contributions, and experimental holo water positions (cyan). For the apo form, the ligand is coloured by the respective interpolated $\Delta_{hyd}G_P$ contributions of replaced apo water molecules (colouring from blue to red from -2.0 to +2.0 in units of kcal/mol). Water positions are shown within 3.5 Å of the ligand. For the 2bq7 holo form, the experimentally determined water position in the S1-pocket is shown in cyan for comparison. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ structures/) in the respective pdb folders.

In Figure 29, a MMP of CKD2 inhibitors is illustrated. The more affine ligand (1pxn:[301] pK_{aff} = 7.15; 1pxp:[301] pK_{aff} = 6.66) bears a methylamino substituent instead of a methyl substituent at the thiazole ring and a phenol instead of a dimethylaniline. Analysis of the *apo* water thermodynamics reveals that the respective binding site contains mostly "happy" water molecules. The interpolated $\Delta_{\text{hyd}}G_{\text{P}}$ contributions show that especially the region where the hydroxyl group of the phenol (or, respectively, the dimethylamino group of the aniline) binds contains highly favourably bound water molecules. The analysis in 4.1 revealed that replacement of such "happy" water molecules by hydroxyl groups correlates with higher ligand affinity, while carbon atoms tend to replace more "unhappy" water molecules in general. Hence, the more affine 1pxn ligand nicely fulfils the derived replacement rules, while a methyl group is located in the respective "happy" water area in the 1pxp complex. Besides, the hydroxyl group likely can undergo better interactions with the nearby Lys sidechain than the dimethylamino group.

Interestingly, for this MMP, also significant differences can be seen when taking into account the *holo* water thermodynamics: In the proximity of the hydroxyl group, all water molecules remain similarly "happy" like in the *apo* form, while the presence of the dimethylamino group leads to the introduction of an unstable hydration site. A similar trend can be observed for the thiazole substituent: In 1pxp, an "unhappy" water molecule can be observed in the proximity of the methyl group that is not observed near the corresponding aminomethyl group in 1pxn. This nicely illustrates that replacement of especially "happy" water molecules by unsuitable ligands groups may not be only disadvantageous because their favourable $\Delta_{\text{hyd}}G_{\text{P}}$ contributions have to be compensated but also because an unfavourable environment for the surrounding water molecules is created, for instance due to the decrease in the number of hydrogen bonding partners.

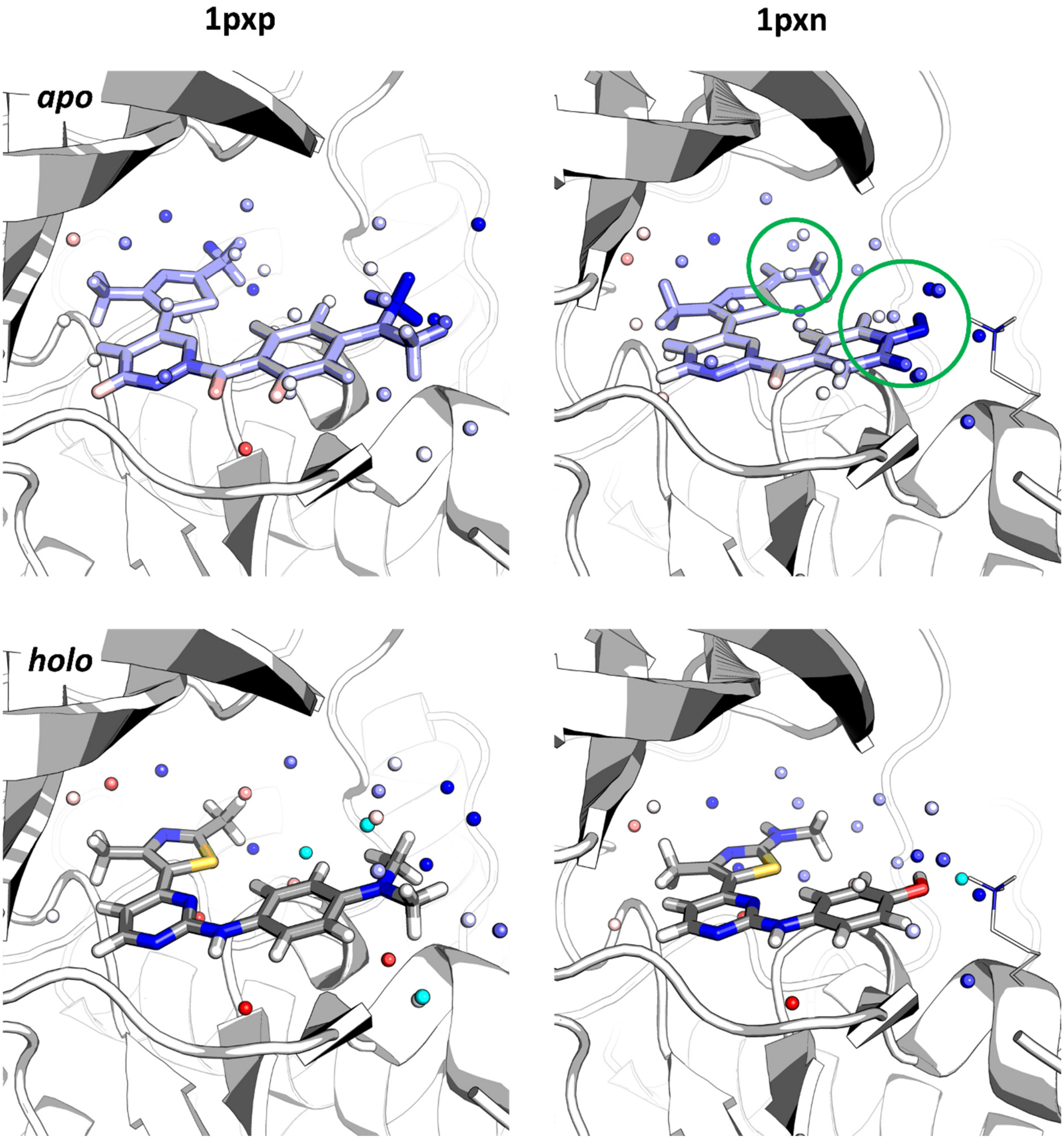

Figure 29: Complex structures 1pxp (pK_{aff} = 6.66) and 1pxn (pK_{aff} = 7.15) with apo and holo water positions as predicted by 3D RISM-based algorithms, coloured by their $\Delta_{hyd}G_P$ contributions, and experimental holo water positions (cyan). For the apo form, the ligand is coloured by the respective interpolated $\Delta_{hyd}G_P$ contributions of replaced apo water molecules (colouring from blue to red from -2.0 to +2.0 in units of kcal/mol). Water positions within 3.5 Å of the ligand are shown. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ structures/) in the respective pdb folders.

Another MMP where substitution of a hydroxyl group plays a role is shown in Figure 30. Formycin A and 5'-methylthiotubercidin are both ligands of *Arabidopsis thaliana* 5'-Methylthioadenosine nucleosidase. 5'-methylthiotubercidin, complexed in 2qtg[302] (pK_{aff} = 5.08), bears a methylsulfanyl moiety instead of an ethylhydroxyl group and contains a different scaffold in the base analogue moiety compared to formycin A, complexed in 2qtt[302] (pK_{aff} = 4.32). Analyses of the *apo* water thermodynamics shows that the binding site region which is accommodating the hydroxyl groups of the sugar ring contains water molecules with highly favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions, thus being nicely in line with the replacement rules derived in 4.1. However, the area where the ethylhydroxyl group binds (or, respectively, the methylsulfanyl group) contains several high energy water molecules. Consequently, the more hydrophobic methylsulfanyl group in 5'-methylthiotubercidin is better suited to replace these water molecules and likely can undergo enhanced interactions with the environment, i.e. the neighbouring Met residue. However, the presence of the methyl group leads to the presence of some “unhappy” water molecules in the *holo* structure, too, so that in this case the affinity gain is probably rather dominated by interactions. Interestingly, when comparing the base analogue moieties, the formycin A scaffold is more in line with the derived replacement rules, with a nitrogen group being located in a “happy” water region. Consequently, there is a high energy water molecule predicted by 3D RISM for the 5'-methylthiotubercidin complex which is not present in the formycin A complex. Thus, when regarding only water thermodynamics, the formycin A scaffold seems to fit the binding site properties more nicely; however, this effect is likely outweighed by the larger impact of the methylsulfanyl substituent. Once again, this MMP thus illustrates how *apo* water properties can be exploited for choosing the optimal substituent at a given ligand.

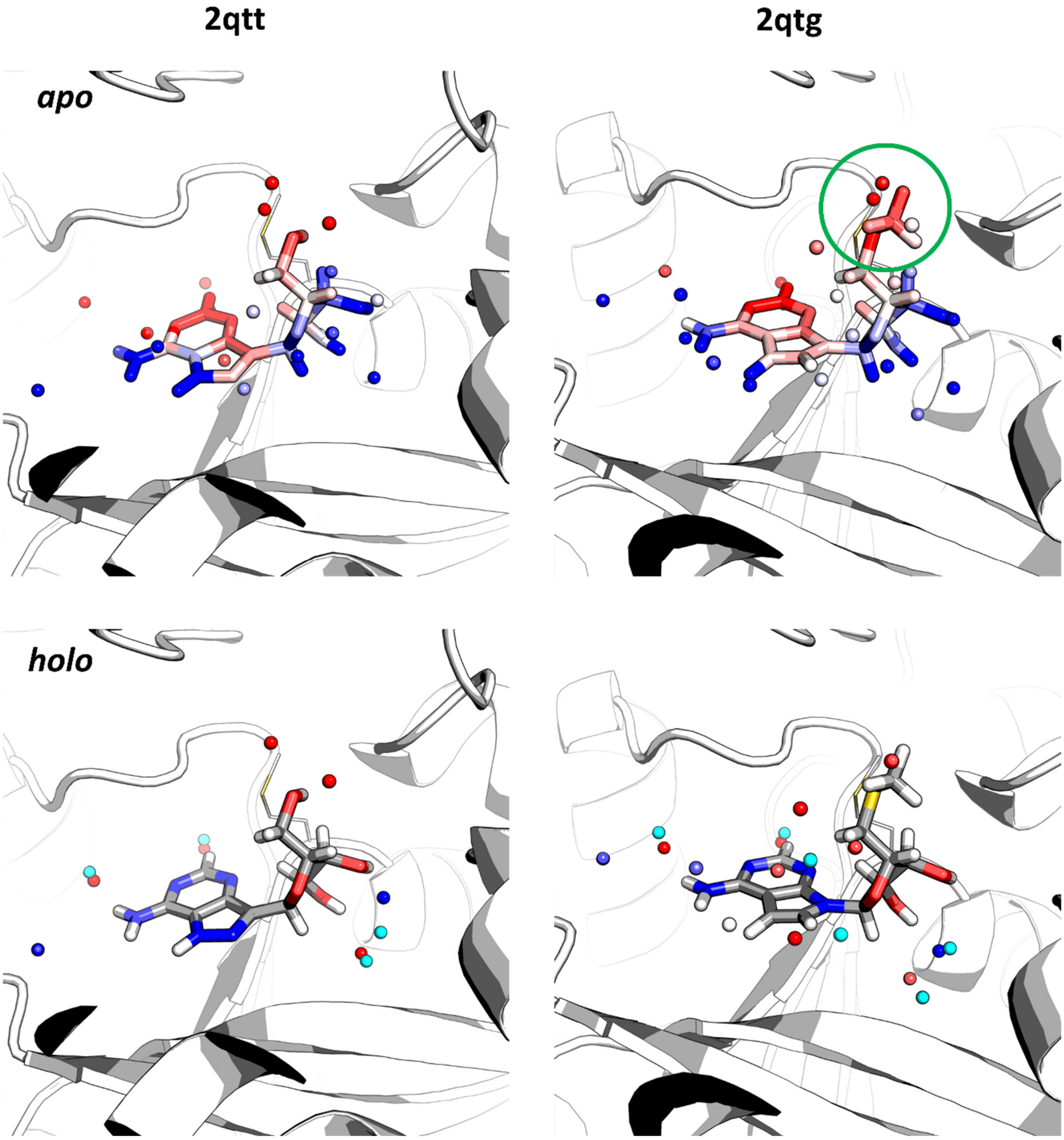

Figure 30: Complex structures 2qtt (pK_{aff} = 4.32) and 2qtg (pK_{aff} = 5.08) with apo and holo water positions as predicted by 3D RISM-based algorithms, coloured by their $\Delta_{hyd}G_P$ contributions, and experimental holo water positions (cyan). For the apo form, the ligand is coloured by the respective interpolated $\Delta_{hyd}G_P$ contributions of replaced apo water molecules (colouring from blue to red from -2.0 to +2.0 in units of kcal/mol). Water positions within 3.5 Å of the ligand are shown. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ structures/) in the respective pdb folders.

4.1.4 Local empirical corrections for 3D RISM

As described in 2.3.3.3, empirical corrections were introduced for the calculation of absolute hydration free energies using 3D RISM that take into account the solute's partial molar volume V_m and charge q and thus make up for known artifacts within the 3D RISM framework. Although the present work is not focused on the calculation of absolute hydration free energies but rather on local contributions of specific spatial regions, it is of interest to investigate if and how the correction affects the local free energy distribution (e.g. if a specific water molecule is considered "unhappy" based on the original $\rho_\mathrm{G}(\mathbf{r})$ field but "happy" based on the corrected field or vice versa).

Therefore, in a proof-of-concept study, the local V_m- and q-based correction was carried out according to Eq. (66) for an exemplary complex within the PDBbind refined set (pdb: 2xbv).

In Figure 31, the respective minima and maxima of the original and corrected $\rho_\mathrm{G}(\mathbf{r})$ fields (i.e. regions with especially favourable and unfavourable contributions to μ^ex or $\mu^\mathrm{ex,corr}$) are shown for two different thresholds. It can be seen that there is no significant visual difference between the two fields – suggesting that the local correction does not affect the local distribution of regions with especially favourable and unfavourable contributions to μ^ex (and thus the localisation of especially "happy" and "unhappy" water molecules).

Yet, an analysis was performed to study the effect of the local correction on the $\Delta_\mathrm{hyd}G_{\mathrm{P},w}$ value of each water molecule: For each water position $\mathbf{r}_w$ as determined from the $g_\mathrm{O}(\mathbf{r})$ field, two respective individual hydration free energy contributions ($\Delta_\mathrm{hyd}G_{\mathrm{P},w,\mathrm{orig}}$ and $\Delta_\mathrm{hyd}G_{\mathrm{P},w,\mathrm{corr}}$) were calculated based on the original and the corrected $\rho_\mathrm{G}(\mathbf{r})$ field. Here, it has to be noted that the current methodical framework did not allow for application of a Gaussian convolution as post-processing, so that the respective "raw" original and corrected $\rho_\mathrm{G}(\mathbf{r})$ fields had to be used. Hence, $\Delta_\mathrm{hyd}G_{\mathrm{P},w,\mathrm{orig}}$ and $\Delta_\mathrm{hyd}G_{\mathrm{P},w,\mathrm{corr}}$ were determined by summation of the $\mu^\mathrm{ex}(\mathbf{r}_i)$ (or $\mu^\mathrm{ex,corr}(\mathbf{r}_i)$) values of all volume elements within 2.5 Å of the water position $\mathbf{r}_w$. This procedure is different from the methodology described in 3.4 and is generally less well suited to determine individual $\Delta_\mathrm{hyd}G_\mathrm{P}$ contributions since the raw fields are highly rugged, so that small deviations in space can lead to large variations in the respective energy values. However, it still allows to estimate the effect of the local PMV correction on the invidual $\Delta_\mathrm{hyd}G_\mathrm{P}$ contributions of the different water molecules.

The scatter plot for the respective values $\Delta_\mathrm{hyd}G_{\mathrm{P},w,\mathrm{orig}}$ and $\Delta_\mathrm{hyd}G_{\mathrm{P},w,\mathrm{corr}}$ based on the original and uncorrected $\rho_\mathrm{G}(\mathbf{r})$ fields is shown in Figure 32. In addition, a visual comparison of the respective binding site with water molecules in the proximity of the ligand coloured according to $\Delta_\mathrm{hyd}G_{\mathrm{P},w,\mathrm{orig}}$ and $\Delta_\mathrm{hyd}G_{\mathrm{P},w,\mathrm{corr}}$ is given in Figure 33.

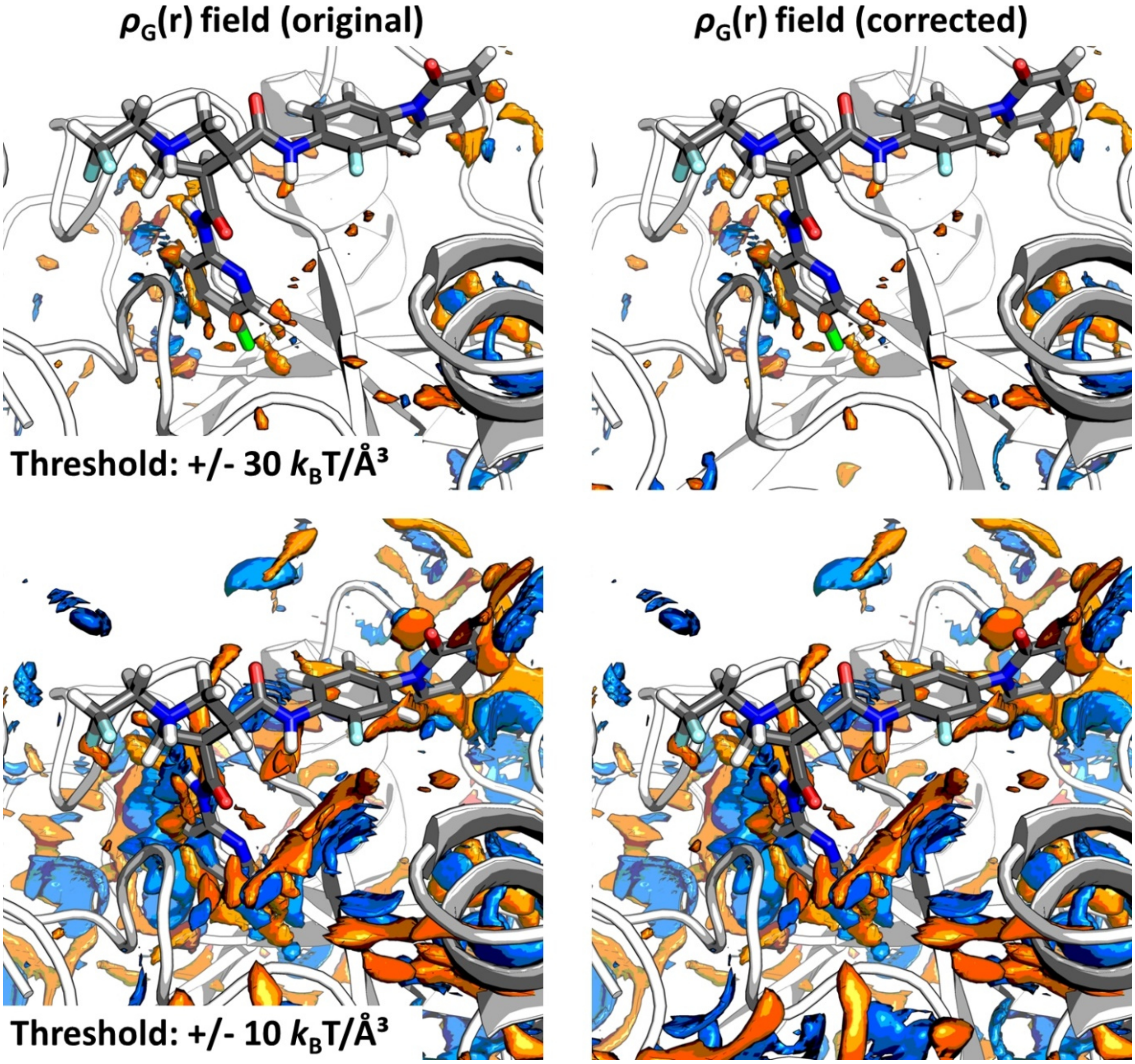
ρ_G(r) field (original)
ρ_G(r) field (corrected)
Threshold: +/- 30 k_BT/Å^3
Threshold: +/- 10 k_BT/Å^3

Figure 31: Minima (blue) and maxima (orange) of the original and corrected $\rho_G(\mathbf{r})$-field (without application of a Gaussian convolution) according to Eq. (66) for the apo structure 2xbv for two different $\rho_G(\mathbf{r})$ thresholds. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/PMV_correction/).

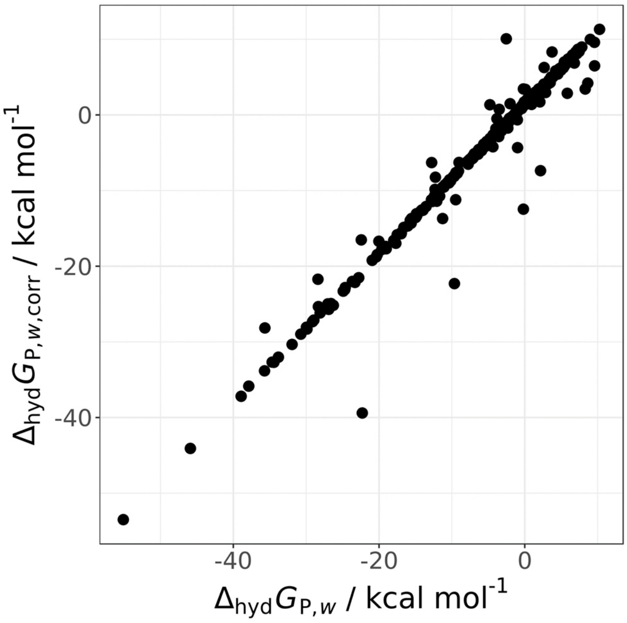

Figure 32: Scatter plot of the original and corrected $\Delta_{hyd}G_{P,w}$ values (in kcal/mol) of predicted water molecules in structure 2xbv. The local correction was carried out according to Eq. (66), and $\Delta_{hyd}G_{P,w}$ values were calculated via summation over all volume elements within 2.5 Å of the water position. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/PMV_correction/).

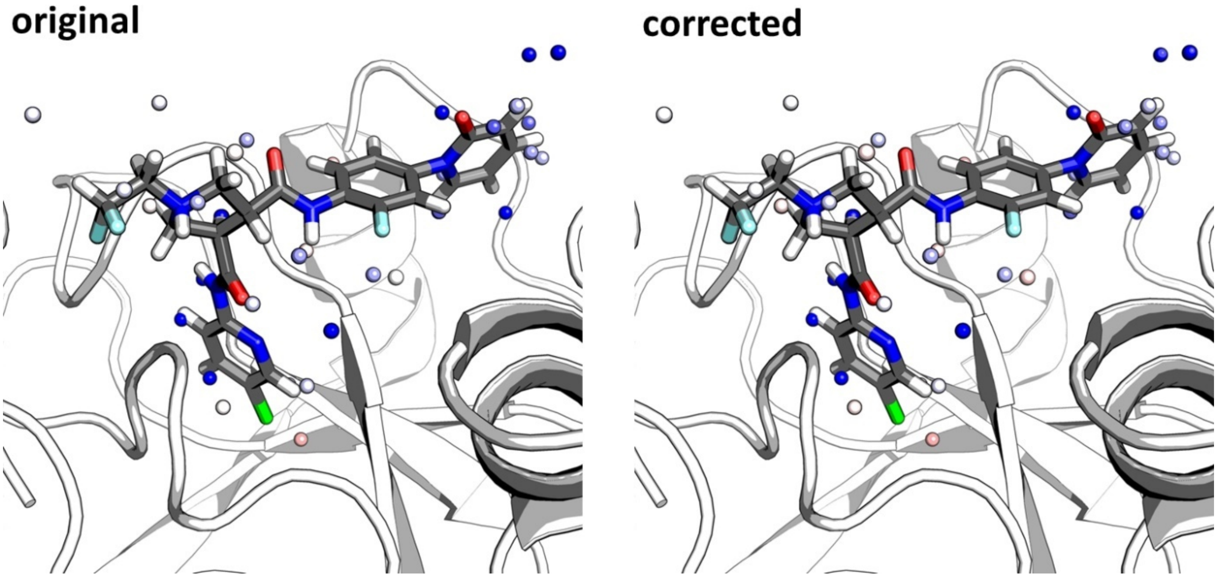

Figure 33: Binding site of structure 2xbv with apo water molecules as predicted by 3D RISM based algorithms within 3.5 Å of the ligand, coloured according to the original and corrected $\Delta_{hyd}G_{P}$ contributions (from blue to red from -17.8 to +17.8 kcal/mol). The local correction was carried out according to Eq. (66), and $\Delta_{hyd}G_{P,w}$ values were calculated via summation over all volume elements within 2.5 Å of the water position. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ PMV_correction/)

Comparison of the original and corrected $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions shows that, with few exceptions, the values are only slightly affected by the correction. The high deviations for some of the water molecules likely result from the fact that the $\rho_{\mathrm{G}}(\mathbf{r})$ fields were not smoothed via a Gaussian convolution in this case, so that certain areas exhibit very high absolute values where small changes result in large absolute shifts. For comparison, the smoothened original $\rho_{\mathrm{G}}(\mathbf{r})$ field, $\rho'_{\mathrm{G}}(\mathbf{r})$, is shown in Figure 34.

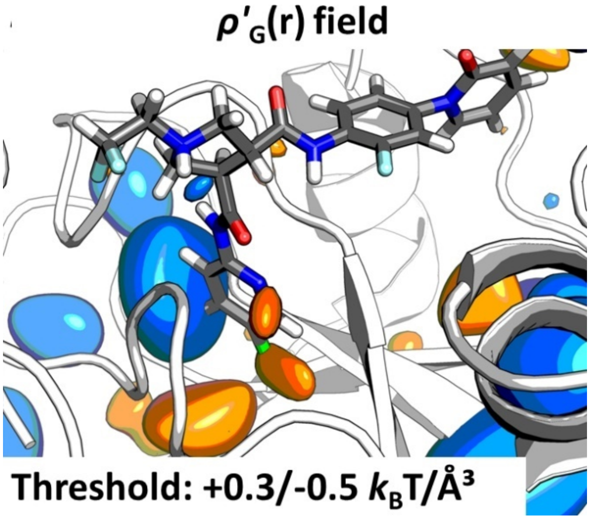

Figure 34: Minima (blue) and maxima (orange) of the $\rho'_{G}(\mathbf{r})$-field obtained by applying a Gaussian convolution with a σ of 1.4 Å to the original $\rho_{G}(\mathbf{r})$ field for the apo structure 2xbv. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/PMV_correction/).

However, in most cases, even a large shift does not result in an inversed trend: for instance, for the water molecule with the largest difference between $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w,\mathrm{orig}}$ and $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w,\mathrm{corr}}$, the value is shifted from -22.3 kcal/mol to -39.4 kcal/mol. In addition, the visual comparison in Figure 33 reveals that no relevant shifts are observed for any of the water molecules in the proximity of the ligand. This is an important finding since it underlines the validity of the presented analysis: Although the V_{m}- and q-based correction is needed for the correct calculation of absolute hydration free energy values, it does not severely affect the local distribution of the hydration free energy, so that the ranking of the water molecules w.r.t. their $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions remains the same (i.e. the most "unhappy" water molecules in the binding site are still the most "unhappy" ones after the correction). A reason for this likely is that the correction does not that much affect the solvent regions but rather the regions where the solute is located.

Although the energetic ordering of the water molecules is not relevantly changed by the correction, and although only small changes are observed for most of the water molecules, it can be interesting to investigate if the correction generally shifts the $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions to more unfavourable or to more favourable values. For the original $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w}$ values, the average is -6.5 ± 11.0 kcal/mol, for the corrected ones -5.2 ± 11.0 kcal/mol, implying only a very slight shift to more unfavourable values (s. also Figure 35). The differences are so small that they would likely be not observable after applying a Gaussian convolution on the respective $\rho_{\mathrm{G}}(\mathbf{r})$ fields.

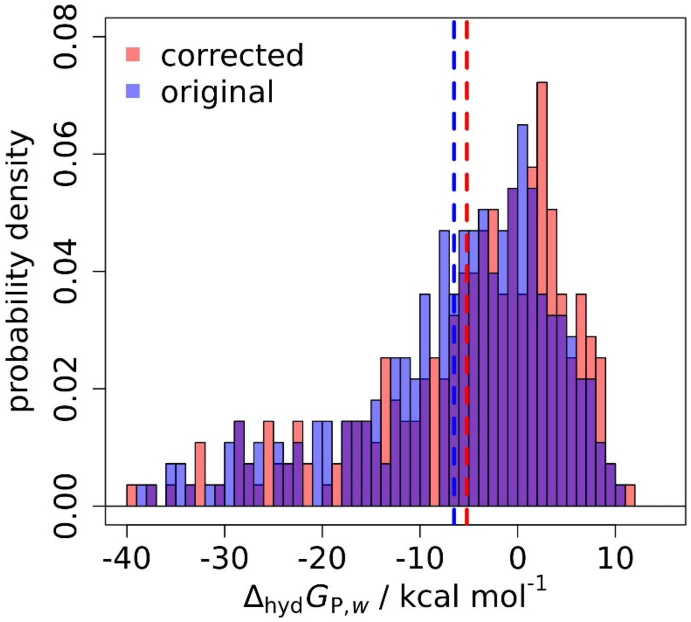

Figure 35: Probability densities of the original and corrected $\Delta_{hyd}G_{P,w}$ values (in kcal/mol) of predicted water molecules in structure 2xbv. The local correction was carried out according to Eq. (66), and $\Delta_{hyd}G_{P,w}$ values were calculated via summation over all volume elements within 2.5 Å of the water position. The average values are shown as dashed lines (blue: original, red: corrected). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_refined_set/ data/ PMV_correction/).

In general, the shift towards negative $\Delta_{\mathrm{hyd}}G_{\mathrm{P},w}$ values for both the corrected and uncorrected values is an interesting finding when considering the results discussed in 4.1.1.5: When comparing the average $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions in the proximity of specific amino acids groups with corresponding results from an analysis based on WaterMap,[278] a highly similar ranking is obtained, but the absolute hydration free energy values are shifted to almost exclusively positive values for the WaterMap calculations. This is a hint that, while both methods are in excellent agreement w.r.t. relative hydration free energy distributions, there might be artifacts in the respective theoretical frameworks that lead to a systematic shift of the absolute values. This might be investigated in the future to improve the calculation of absolute $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions. In this context, a direct comparison of results from 3D RISM with other methods, like WaterMap, on the very same protein structures could be highly beneficial.

4.2 Binding site characterisation based on 3D RISM *uu* calculations

In the previous chapters, the focus was set on the local hydration site thermodynamics within *apo* and *holo* binding sites and their correlation with ligand features. In this chapter, the concept of local thermodynamic characterisation will be extended towards a more general thermodynamic signature of binding sites: Inspired by Goodford's GRID approach, 3D RISM solute-solute (*uu*) calculations were carried out to determine the distribution of specific probes, mimicking ligand functional groups, within binding sites. This allows to get a detailed thermodynamic binding site profile considering both hydration and the distribution of distinct pharmacophore features. Similar to the $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions and the solvent site densities, the probe densities can be interpolated and mapped to the atoms of a bound ligand, e.g. from an experimental complex or from docking. Thus, the obtained probe densities can be exploited for the *de novo* design of ligands that match the binding site profile, or to complement scoring in virtual screening. In the following chapters, the concept will first be validated by analysing the correlation between probe densities and ligand features based on the PDBbind core set 2013. Afterwards, a ligand-probe matchscore will be introduced that will be validated for both pose recovery and virtual screening. Finally, a workflow will be presented to exploit the probe densities for the fragment-based *de novo* design of novel ligands.

All raw data for the analyses in 4.2 can be found in the Electronic Appendix (Electronic_Appendix/ PDBbind_core_set/ and Electronic_Appendix/XIAP/). This includes respective ligand and protein structures (ligand.pdb, pocket.pdb) as well as the interpolated probe *g*-function data on the ligands (gUU_xy@lig.pdb) for each structure within the PDBbind core set and for the discussed XIAP structures with respective ligands and docking poses.

4.2.1 Correlation between probe densities and native ligands

In the presented work, three simplistic, spherical probes derived from GAFF atom types (s. 3.2) were employed to mimic distinct pharmacophoric features, namely an uncharged c3 probe as well as a positively charged n4 and a negatively charged o probe. Using the 3D RISM *uu* formalism, respective density fields of each probe within an *apo* binding site binding site are obtained which can be used for ligand design.

However, to validate the concept of 3D RISM *uu*-derived pharmacophoric probe densities, it is first necessary to correlate the different densities within a binding site with the structure of known ligands. Under the assumption that a bound ligand shows near-ideal interactions with the protein, peaks for a specific pharmacophoric probe should coincide with corresponding ligand atoms.

An illustrative example of the probe densities and their matching with ligand features is presented in Figure 36. It shows the binding site of XIAP (pdb: 5c7a[261]) with respective densities of the three used probes in the respective *apo* binding site and their interpolation onto the bound ligand.

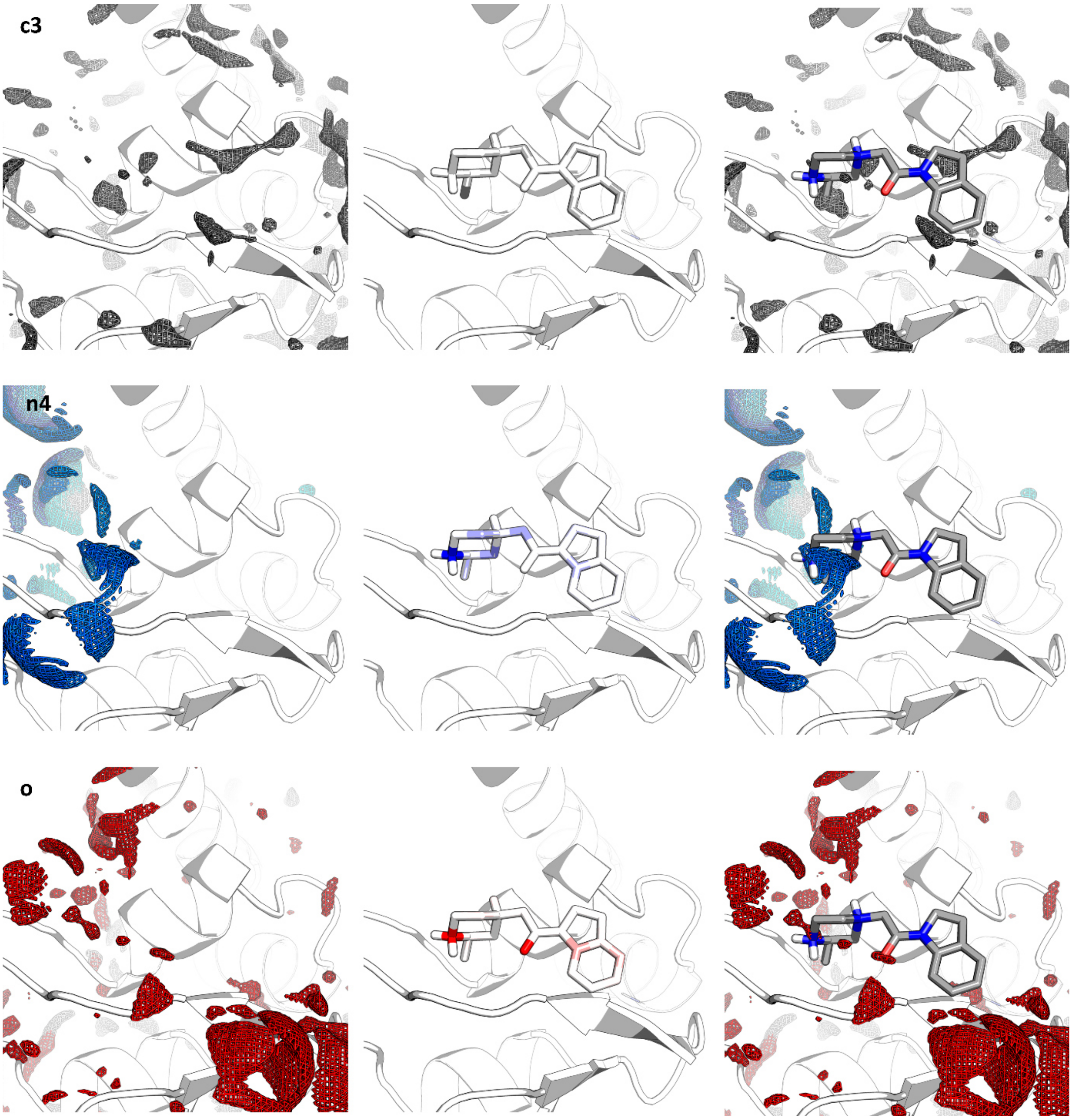

Figure 36: Probe densities and interpolated densities mapped onto ligand atoms for 5c7a. Upper row: c3 probe (density threshold = 10, colouring from white to grey from 0 to 8), middle row: n4 probe (density threshold = 300, colouring from white to blue from 0 to 50), lower row: o probe (density threshold = 30, colouring from white to red from 0 to 10). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/ 5c7a/).

While the density fields (left column) themselves seem rather abstract, comparison of the resulting mapping and the ligand structure (middle and right column) reveals a striking matching: The amine group and the carbonyl group in the ligand coincide with respective probe density peaks. Interestingly, the region around the nitrogen atom in the piperazine ring in the western region of the binding site shows a high density for both the n4 and o probe, implying that this area could potentially accommodate various polar groups. Carbon atoms are naturally more abundant in small molecules; yet, very good agreement can also be observed for the c3 probe: Especially the methyl group at the piperazine ring and carbon atoms of the indole ring exhibit high interpolated c3 probe density values.

To quantify this finding, respective 3D RISM *uu* calculations were performed for the *apo* proteins of the complexes in the PDBBind core set 2013, and the interpolated probe *g*-function values on the given ligand atoms *l* was determined. The results for the respective average probe *g*-function values at i) atoms with matching atom type, ii) atoms with a matching element, and iii) all other atoms are shown in Table 20. In case of the c3 probe, also atoms of the ca type were considered a matching atom type, and since it is meant to be a more general probe for apolar groups, halogen atoms were also considered a matching element.

Table 20: Average interpolated apo probe g-function values $g_{l,p}$ with standard deviation of the n4, o and c3 probe at ligand atoms of i) the corresponding type (n4, o, and c3 and ca for the c3 probe), ii) the corresponding element (N, O, C/Halogen), and iii) at all other atoms on respective ligand atoms for all structures in the PDBbind core set. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_core_set/ data/ Table20_uu_mapping/).

	$g_{l,\mathrm{n4}}$	$g_{l,\mathrm{o}}$	$g_{l,\mathrm{c3}}$
corresponding atom type	48.1 ± 46.6	35.8 ± 43.8	3.7 ± 8.7
corresponding element	29.1 ± 41.5	28.1 ± 41.4	4.1 ± 10.4
other atoms	17.3 ± 34.2	5.9 ± 19.4	3.1 ± 9.6

The results in Table 20 show strong trends for the n4 and o probe, with ligand atoms of a matching type having the highest average interpolated probe *g*-function values, followed by atoms of a matching element, while all other atoms exhibit lower values, especially for the o probe. The less pronounced trend for the c3 probe meets expectations since carbon atoms are much more abundant in ligands so that naturally many carbon atoms are located in areas with lower probe density. Nevertheless, higher values are observed for atoms of the respective type and element than for other atoms. To further investigate if areas with high c3 probe density coincide with apolar regions in the ligand, the average interpolated c3 probe *g*-function value was determined on carbon atoms with no polar atoms (nitrogen, oxygen, sulphur, or phosphorous) within 3.0 Å. Indeed, the resulting value (11.0 with a high standard deviation of 25.6) is much higher, suggesting that hydrophobic parts of the ligands match areas of high c3 probe density. The generally lower values for the c3 probe compared to the charged n4 and o probes can be simply attributed to the lack of electrostatic interactions for this probe.

Since the three probe atoms have different charges, it was also analysed if there is a correlation between the partial charge of a given ligand atom and the respective interpolated probe *g*-function values. In Table 21, the average interpolated *g*-function values of the n4, o and c3 probes are shown for subsets of ligand atoms with different partial charges.

Table 21: Average interpolated apo probe g-function values $g_{l,p}$ with standard deviation of the n4, o and c3 probe at atoms with different partial charges q. for all ligand atoms in structures in the PDBbind core set. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ PDBbind_core_set/ data/ Table21_probe_values_partial_charges/).

q	$g_{l,\mathrm{n4}}$	$g_{l,\mathrm{o}}$	$g_{l,\mathrm{c3}}$
$q > 0.75$	6.1 ± 22.9	3.6 ± 15.3	1.8 ± 3.1
$q > 0.5$	12.0 ± 29.7	3.2 ± 12.8	2.2 ± 6.3
$q > 0.25$	12.2 ± 30.4	4.0 ± 15.6	1.6 ± 4.8
$q > 0.1$	15.4 ± 32.7	4.4 ± 16.5	2.8 ± 8.3
$-0.1 < q \leq 0.1$	19.8 ± 36.2	6.8 ± 20.5	4.0 ± 11.4
$q < -0.1$	18.6 ± 35.1	14.7 ± 32.2	3.6 ± 9.5
$q < -0.25$	17.5 ± 34.8	22.4 ± 38.0	2.4 ± 7.6
$q < -0.5$	16.8 ± 34.9	26.4 ± 40.3	2.3 ± 8.0
$q < -0.75$	14.0 ± 32.3	39.3 ± 44.9	1.7 ± 5.5

Indeed, there is a clear trend for the o probe: The average interpolated g_o value significantly increases with decreasing partial charge of the ligand atom, i.e. negatively charged atoms in average have significantly higher interpolated g_o values. This meets expectations since peaks of the negatively charged o probe are likely found in an environment suitable for accommodating negatively charges groups like carboxylates. For the n4 probe, an opposite trend could be expected – however, no trend w.r.t. the partial charge subsets is observed here. The reason for this simply is that, although groups like tertiary amines have a net positive charge, the partial charges of the nitrogen and hydrogen atoms within these groups have opposite signs. Therefore, when an n4 probe density peak overlaps with the position of e.g. a ternary amine, high interpolated n4 probe densities are observed on atoms with negative and positive partial charges alike. Hence, no significant correlation between atom partial charges and interpolated g_n4 is seen here.

For the uncharged c3 probe, on the other hand, the highest interpolated g_{c3} values are observed for subsets of ligand atoms with the lowest absolute charges. This again meets expectations since c3 probe density peaks are likely found in binding site areas with few charged residues which are suitable to accommodate uncharged ligand moieties. Hence, the correlation of the average, interpolated probe *g*-function values and the atom partial charges is in line with the results in Table 20 and emphasises the relevance of the charge parameter for the probe distribution.

All in all, the presented analyses thus confirm that the probe densities are highly correlated with the ligand atoms' element identities and properties thus can be exploited for assessing and designing ligand structures.

4.2.2 Introducing a score for ligand-probe-matching

To achieve a quantitative measure for a ligand structure-probe density match, a respective score was developed with the aim to capture how well the ligand atoms are in line the probe fields, e.g. if nitrogen atoms are located at positions with high *g*-function values of the n4 probe.

To quantify this, all atoms *l* of a molecule *m* with an interpolated probe *g*-function value $g_{l,p}$ exceeding a certain threshold t_p (here: 20 for n4, 10 for o and 3.5 for c3, derived from the analysis on the PDBbind core set) are considered maxima in the respective probe fields, resulting in *N* maxima of a given probe *p* in a molecule structure *m*, depending on threshold t_p, $N_{\text{max},p,m}(t_p)$.

$$N_{\max,p,m}(t_p) = \left|\left\{ j \in l \mid g_{j,p} > t_p \right\}\right| \tag{73}$$

For each ligand atom, it is evaluated if this atom (or a neighbouring atom within a distance *d* of 1.5 Å) has the "right" corresponding element type *e*, i.e. if the maximum is matched or not, resulting in *N* probe maxima in a ligand that are considered "fulfilled", $N_{\text{match},p,m}(t_p)$:

$$N_{\text{match},p,m}(t_p) = \left|\left\{ j \in l \mid g_{j,p} > t_p \text{ and } (e_j = e_p \text{ or } e_k = e_p, k \in l \mid d(k,j) <= 1.5 \text{ Å}) \right\}\right| \tag{74}$$

Thus, for each molecule *m*, a ratio $x_{\text{fulfilled\_peaks},p,m}(t)$ can be determined for each probe *p* as the ratio of $N_{\text{match},p,m}(t_p)/N_{\text{max},p,m}(t_p)$:

$$x_{\text{fulfilled\_peaks},p,m}(t_p) = \frac{N_{\text{match},p,m}(t_p)}{N_{\max,p,m}(t_p)} \tag{75}$$

Thus, $x_{\text{fulfilled\_peaks},p,m}(t_p)$ has a value of 1.0 if all maxima coincide with a "correct" element and 0.0 if none of them do.

Additionally, it is also important to account for mismatches. The ratio $x_{\text{fulfilled\_peaks},p,m}(t_p)$ only captures if the "correct" atom is present in an area with high density for a given probe. It does not capture if there are other atoms of the respective type within the ligand which lie in areas with low probe density. However, it should also be considered that atoms with low interpolated density values should NOT be an element of the respective type. Therefore, another measure, termed $x_{\text{fulfilled\_elements},p,m}(t_p)$, is introduced as the ratio of $N_{\text{match},p,m}(t_p)$ and the total number of atoms of the corresponding element type in the ligand, $N_{p,m}$:

$$x_{\text{fulfilled\_elements},p,m}(t_p) = \frac{N_{match,p,m}(t_p)}{N_{p,m}} \tag{76}$$

Thus, $x_{\text{fulfilled\_elements},p,m}(t_p)$ has a value of 1.0 if all atoms of a given element are considered peaks of the respective probe, for instance if all oxygen atoms within the molecule have interpolated g_o values exceeding the chosen threshold, and 0.0 if none of them are peaks. $x_{\text{fulfilled\_peaks},p,m}(t_p)$ and $x_{\text{fulfilled\_elements},p,m}(t_p)$ are then multiplied to get an overall match score for each probe p, $x_{p,m}(t_p)$, which is zero if $x_{\text{fulfilled\_peaks},p,m}(t_p)$ or $x_{\text{fulfilled\_elements},p,m}(t_p)$ is zero.

$$x_{p,m}(t_p) = x_{\text{fulfilled\_peaks},p,m}(t_p) \cdot x_{\text{fulfilled\_elements},p,m}(t_p) \tag{77}$$

A total score for the binding mode of molecule *m* (or, in case of docking, a molecule pose), *s*, is then obtained via summation of all probe scores (divided by the number of probes N_p, in this case 3, to obtain a range from 0.0 to 1.0):

$$s_m(\{t_p\}) = \frac{\sum_p x_{p,m}(t_p)}{N_p} \tag{78}$$

Alternatively, if the score has to be more restrictive, for instance for virtual screening purposes, when a lot of poses should be filtered out, it can also be obtained as the product of all probe scores, resulting in a total score of zero if one of the probe scores is zero:

$$s_m^*(\{t_p\}) = \prod_p x_{p,m}(t_p) \tag{79}$$

It is important to consider that both resulting scores are derived from ratios and are thus designed to not depend on molecule size. This was chosen since the intention is not do build a full scoring function but to obtain a measure for how well a molecule matches the RISM-derived thermodynamic binding site profile. Especially, it is intended to employ it for the *de novo* design of novel ligands via virtual fragment screening and for guiding the design of target-focused combinatorial libraries. For this purpose, it is vital not to find the molecules with the already highest affinity but the best matching fragments which can then be combined into a larger and more affine ligand.

4.2.3 Pose recovery of native ligands

As the analysis based on the PDBbind core set 2013 confirmed the correlation between pharmacophoric probe peaks and the ligands' element identity, a probe-ligand matchscore was introduced to capture how well a given ligand binding pose matches the thermodynamic binding site profile. As described above, it evaluates i) if probe maxima coincide with the presence of respective elements, and ii) if, vice versa, the presence of respective atoms in a ligand coincides with "correct" corresponding probe *g*-function maxima. Thus, it is a measure for how perfectly a given ligand conformation fits the thermodynamic signature of the protein.

To assess the utility of the ligand-probe matchscore for the assessment of docking results, a redocking of the ligands in the PDBbind core set 2013 was performed. For each ligand, three similar docking runs were carried out to create 100 diverse (inter-pose RMSD > 1.5 Å) solutions each time, resulting in a total of 300 poses per ligand. For all of them, the ligand-probe matchscore, s_{pose}, was calculated. In Table 22, the average s_{pose} values for pose subsets based on RMSD over the whole data set are shown. In addition, also the average RMSD values for pose subsets based on s_{pose} were determined; they are given in

Table 23.

Table 22: Average values of s_{pose} as defined in Eq. (78) with standard deviations for docking pose subsets based on RMSD for the docking poses of ligands in the PDBBind core set 2013 (300 diverse poses per ligand). The respective raw data can be found in the Electronic Appendix (RMSD values and s_{pose} values: Electronic Appendix/ PDBbind_core_set/ data/ Table22_Table23_score pose recovery/; docking poses with interpolated probe g-function values $g_{l,p}$ are given in the respective pdb folders in Electronic Appendix/ PDBbind_core_set/ structures/).

RMSD / Å	s_{pose}
< 1.0	0.254 ± 0.076
1.0 - 1.5	0.250 ± 0.114
1.5 - 2.0	0.223 ± 0.124
2.0 - 2.5	0.209 ± 0.122
2.5 - 3.0	0.208 ± 0.118
3.0 - 3.5	0.184 ± 0.111
3.5 - 5.0	0.172 ± 0.101
> 5.0	0.156 ± 0.089

Table 23: Average RMSD values with standard deviations for docking pose subsets based on s_{pose} as defined in Eq. (78) for docking poses of ligands in the PDBBind core set 2013 (300 diverse poses per ligand). The respective raw data can be found in the Electronic Appendix (RMSD values and s_{pose} values: Electronic Appendix/ PDBbind_core_set/ data/ Table22_Table23_score pose recovery/; docking poses with interpolated probe g-function values $g_{l,p}$ are given in the respective pdb folders in Electronic Appendix/ PDBbind_core_set/ structures/).

s_{pose}	RMSD / Å
$s_{\text{pose}} > 1/3$	3.7 ± 2.4
$1/4 < s_{\text{pose}} < 1/3$	4.3 ± 2.7
$1/6 < s_{\text{pose}} < 1/4$	5.5 ± 2.9
$1/12 < s_{\text{pose}} < 1/6$	6.2 ± 2.8
$s_{\text{pose}} < 1/12$	6.0 ± 2.7

The results show a clear trend of higher s_{pose} values for poses with lower RMSDs to the native binding mode; correspondingly, the average RMSD values for subsets of poses with higher ligand-probe matching scores are lower. Usually, poses with an RMSD threshold of 2.5 Å are considered satisfactory. Interestingly, even within these satisfactory poses, a trend for higher average s_{pose} values for poses with especially low RMSD values below 1.5 and 1.0 Å can be observed. This implies that the (near) native binding modes exhibit the best matching with the probe densities and that the s_{pose} score can be used to distinguish between “good” and implausible ligand poses. This is quite noteworthy since - while the interaction energy of the probes is included 3D RISM *uu* formalism - the score does not include any information about the ligand’s conformation’s internal energy. The results thus underline that the probe densities include valuable information about the binding site thermodynamics and can be exploited for SBDD purposes.

Encouraged by these results, it was analysed if maybe even a direct correlation could be observed between the probe-ligand matchscore of the bound ligand conformation and the ligand affinity within the PDBbind core set 2013. A respective correlation analysis revealed that this is not the case (R value: -0.027, p value: 0.12; affinity and score data can be found in the Appendix, 7.6). However, this can likely be attributed to the heterogeneity of the binding sites within the data set which comprises several different protein families. Due to the different sizes and characteristics of the binding sites of different proteins, conventional docking scores are usually not comparable for different proteins and hence do not allow for a direct conversion of scores to actual binding affinities (s. 2.2.2.2.4).

4.2.4 Virtual screening on XIAP

Encouraged by the good results from the pose recovery docking experiments, the next goal was to probe the usefulness of the probe densities for an even more challenging task, virtual screening. Therefore, docking of the XIAP benchmarking data set provided by DUD-E was performed which comprises 129 active and 5213 corresponding decoy molecules. As already mentioned, the ligand-probe matching score is not designed to be a complete scoring function as it completely neglects any internal ligand energy. Besides, it does not correlate with the molecule size, which is the case for most scoring functions (like the GOLD ChemPLP scoring function which was used here), and which is by trend beneficial for classical virtual screening benchmarks since usually larger molecules indeed have higher affinity. Therefore, it cannot be expected that the ligand-probe matching score alone can compete with the scoring quality of an approved and well-established scoring function like ChemPLP. Rather, the aim was to investigate whether it can be used as a filtering criterion to further improve the enrichment of active molecules by sorting out molecules with high scores but a bad matching with the thermodynamic binding site profile.

Therefore, different combinations of the ligand-probe matching score and ChemPLP were probed for filtering down poses obtained by docking, and respective ROC AUC values were determined as a measure for the differentiation of active and decoy molecules (Table 24). A detailed description of the ROC AUC measure can be found in 2.2.2.2.4.

AUC values can range from 0 to 1, with 0.5 denoting a random ranking of active and decoys and 1.0 a perfect differentiation for which all actives are ranked before all decoys. The results for the XIAP benchmark data set (Table 24) show that GOLD’s ChemPLP scoring function already yields a very good differentiation with an AUC of 0.81 (Figure 37). Intriguingly, when scoring the molecules only based on the ligand-probe matching scores, s_{pose} and the more restrictive s_{pose}^{*}, AUC values of 0.72 are obtained. This is not as high as for ChemPLP, yet surprisingly good when considering the simplicity of the measure, once again highlighting the validity und utility of the concept.

Besides, the AUC of the s_{pose}- and s_{pose}^{*}-based scoring was determined after filtering out molecules with low ChemPLP scores < 60, 70, and 75, respectively. In case of the less restrictive s_{pose}, this does not result in an improved AUC while a considerable increase can be observed for s_{pose}^{*} to an AUC of 0.77, 0.83, and 0.85, respectively.

Table 24: ROC AUC values for the scoring of poses obtained by docking of the DUD-E benchmark data set for XIAP by ChemPLP score, the ligand-probe matching scores s_{pose} and s_{pose} as defined in Eqs. (78) and (79) and combination of them for filtering purposes. For those cases where a filtering is applied (i.e. only docking poses with a ChemPLP or ligand-probe matching score above a certain threshold are retained for the subsequent scoring), the number of remaining molecules after filtering is given in the 3rd column; x_o and x_{n4} denote the matching with only o and n4 probe as defined in Eq. (77). The respective ROC curves can be found in the Appendix (7.7). The respective raw data (poses, scores, interpolated probe g-function values) can be found in the Electronic Appendix (Electronic Appendix/ XIAP/ data/ DUD-E_VS /).*

Score	AUC	*N*(molecules)
ChemPLP	0.81	5342
s_{pose}	0.72	5342
s_{pose} (ChemPLP > 60)	0.75	2648
s_{pose} (ChemPLP > 70)	0.72	491
s_{pose} (ChemPLP > 75)	0.74	178
ChemPLP ($s_{\mathrm{pose}} > 1/6$)	0.86	2580
ChemPLP ($s_{\mathrm{pose}} > 1/4$)	0.95	778
ChemPLP ($s_{\mathrm{pose}} > 1/3$)	0.99	199
ChemPLP (x_{o} and $x_{\mathrm{n4}} > 0.0$)	0.97	612
ChemPLP (x_{o} or $x_{\mathrm{n}} > 0.0$)	0.86	3356
ChemPLP ($x_{\mathrm{n}} > 0.0$)	0.93	2397
ChemPLP ($x_{\mathrm{o}} > 0.0$)	0.87	1571
s_{pose}^*	0.72	5342
s_{pose}^* (ChemPLP > 60)	0.77	2648
s_{pose}^* (ChemPLP > 70)	0.83	491
s_{pose}^* (ChemPLP > 75)	0.85	178
ChemPLP ($s_{\mathrm{pose}}^* > 0.0$)	0.97	609

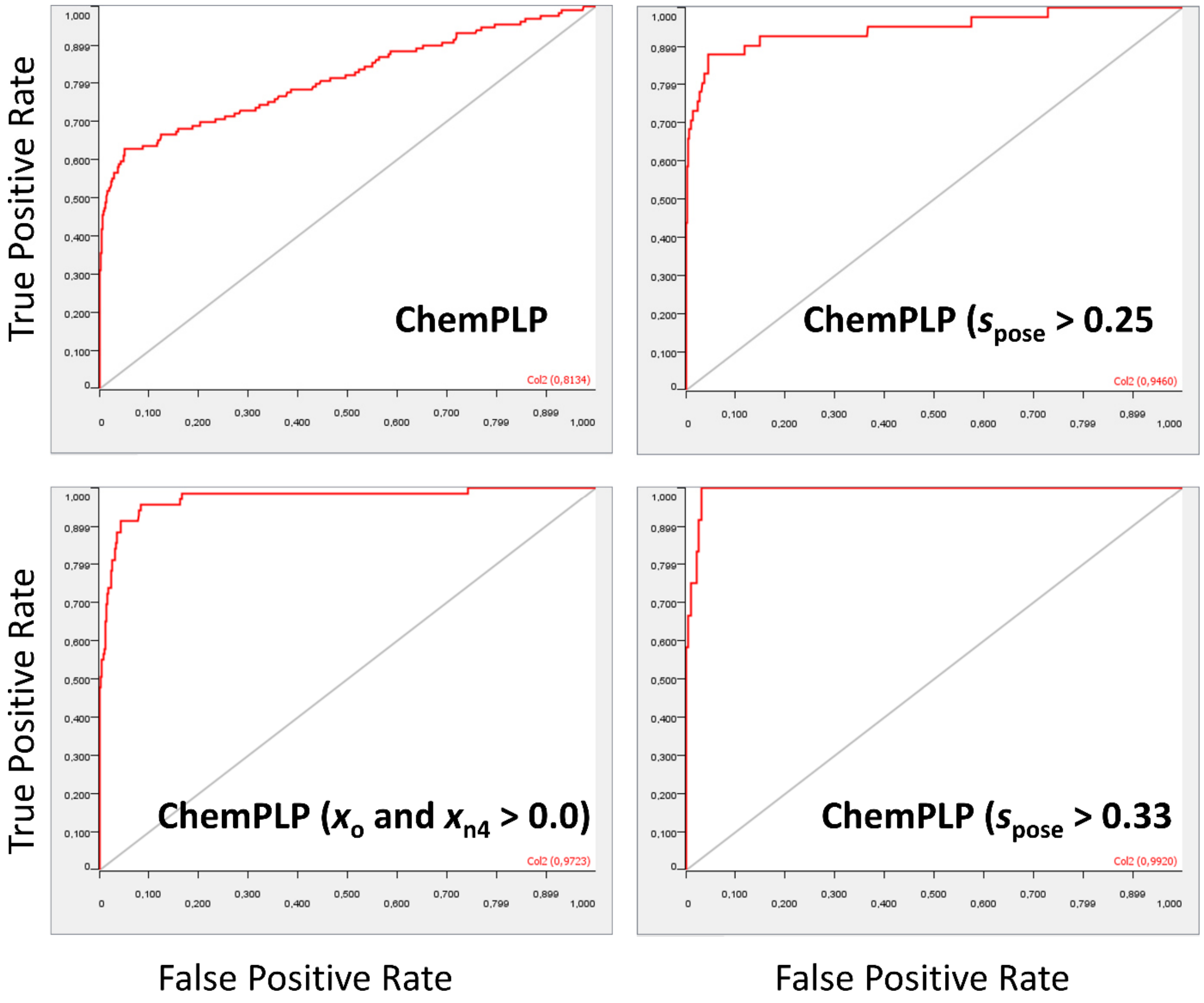

Figure 37: Selected ROC curves for the scoring of poses obtained by docking of the DUD-E benchmark data set for XIAP by ChemPLP score, the ligand-probe matching scores s_{pose} and s_{pose} as defined in Eqs. (76) and (77) and combination of them for filtering purposes. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/ data/ DUD-E_VS/). All ROC curves are given in the appendix (7.7).*

The most promising strategy for combining the ligand-probe matching score with classical scoring, however, is to filter the poses according to the ligand-probe matching score and to then perform ranking w.r.t. the ChemPLP score of the remaining poses. Filtering out all poses with s_{pose} <= 1/6, 1/4, and 1/3 (Figure 37) leads to a significant increase of the AUC to 0.86, 0.95, and 0.99, respectively, i.e. a lot more decoys than actives were filtered out successfully. In case of the > 1/3 threshold, the AUC of 0.99 implies a nearly perfect ranking of the remaining molecules, which is extremely valuable if one wants to select a small number of compounds for experimental testing. Large improvements in the AUC can also by obtained when considering only those poses for which the ligand-probe matching score for either the n4 probe, the o probe, the n4 or the o probe, or the n4 and the o probe (Figure 37) is > 0 (resulting AUCs: 0.93, 0.87, 0.86, 0.97), with again a nearly perfect ranking when both the o and n4 density are matched. The considerable increase to 0.93 when including all poses with matching for the n4 probe implies an especially high relevance of this probe for the binding site thermodynamics. This is in agreement with literature since the region with highest n4 probe density, the P1 pocket of the XIAP binding site which accommodates the piperazine ring in 5c7a, was shown to be highly important for ligand binding affinity.[303]

The s_{pose}^{*} score by definition is zero if one of the probe densities is not matched. Filtering by s_{pose}^{*} thus is similarly restrictive as keeping only poses with $s_{\mathrm{pose}} > 0$ for the o and n probe and, too, yields a nearly perfect ranking of the remaining molecules with an AUC of 0.97.

The results thus show that combination of classical scoring with the ligand-probe matching score can be successfully used to filter out molecules which have relatively high ChemPLP scores but do not match the thermodynamic binding site profile well, resulting in a massive improvement in ranking power for the remaining molecules. It is therefore advocated to use the ligand-probe matching score as a filter criterion for the post processing of docking results in case of virtual screening, with the threshold depending on the desired level of restrictiveness.

4.2.5 Virtual fragment library screening on XIAP

As described, the intended purpose of the ligand-probe matching score is to find molecular fragments which match the thermodynamic binding site profile of a given target and can be selected for the design of novel, high affinity ligands. As a first benchmark study, the fragment library as described by Sandór *et al.*,[260] plus three fragments derived from the fragment-like ligand in 5c7a (Figure 38), were docked into the same XIAP structure as used for the virtual screening benchmark.

Fragment 1 only comprises the piperazine ring and the carbonyl group. Fragment 2 directly corresponds to the ligand in 5c7a. For fragment 3, only the methyl substituent of the piperazine ring was removed to analyse to what extent small structural changes influence the scoring. The aim was to retrieve these three fragments with 10, 18, and 19 heavy atoms from the other 189 fragments, whose heavy atom count ranges from 6 to 22.

A prerequisite for the ranking of fragments is the correct positioning by the docking program. In Figure 38, the top docking poses of the three fragments in 3hl5 are depicted in overlay with the complex structure 5c7a (ligand in grey, protein in white). The overlay shows that there are no considerable structural discrepancies between 3hl5 (pale green) and 5c7a (white) in the respective binding site region, and that the docking poses are in excellent agreement with the overlaid crystal structure in case of fragments 2 and 3. For the smaller fragment 1, the overlap with the native binding mode of the ligand in 5c7a is not as perfect but still sufficiently good considering that the binding site was defined as a rather large region of 10 Å radius around the native ligand in 3hl5.

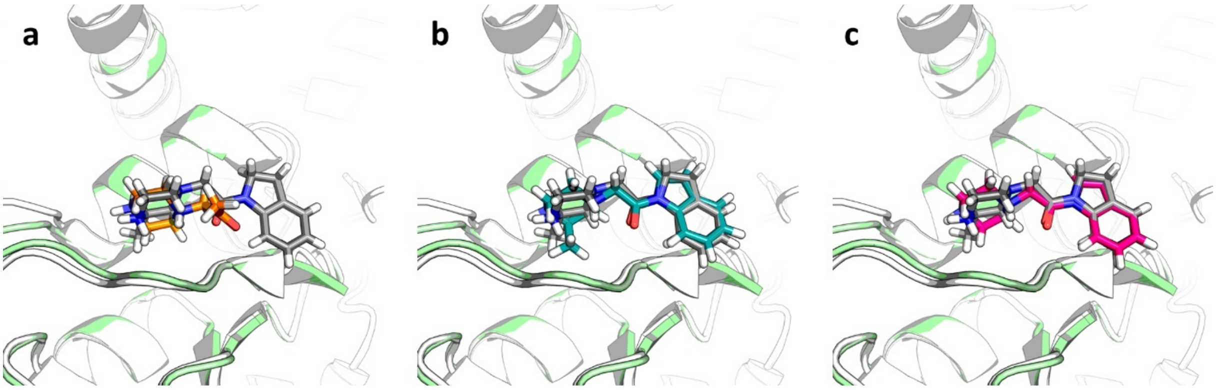

Figure 38: Top ranked docking poses of the three fragments derived from 5c7a by ChemPLP as generated by GOLD in 3hl5 (pale green) in overlay with the complex structure 5c7a (protein: white, ligand: grey). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/ fragments_screening/ (fragment 1:pose 104_5; fragment 2: pose 103_5; fragment 3: pose 101_6)).The respective ranks of the three fragments using conventional scoring and different combinations with the ligand-probe matching score are presented in Table 25.

Different than for classical virtual screening, the dependence of classical scoring functions like ChemPLP on molecule size can be problematic for fragment screening since larger fragments will *per se* obtain higher scores, which hampers the search for small but promising novel scaffolds that show ideal agreement with the thermodynamic profile of the respective binding site region.

Indeed, when ranking only according to ChemPLP, the respective fragments 1, 2, and 3 are found on ranks 71, 1, and 4, respectively (Table 25). While the quite large fragments 2 and 3 are among the top ranked molecules, the smaller fragment 1 could not be retrieved via conventional scoring.

To account for molecule size in virtual screening, it is a common practise to normalise docking scores by molecule size, for instance heavy atom count. When performing ranking according to ChemPLP score divided by the heavy atom count, the resulting ranks of fragments 1, 2, and 3 are 1, 41, and 32 (Table 25). Thus, employing this modified scoring scheme, the smallest fragment could be retrieved, while the two larger ones are no longer among the top ranked molecules, which is again not satisfactory.

Table 25: Ranks of fragments 1, 2, and 3 in the docked fragment data set for scoring according to ChemPLP, ChemPLP normed by heavy atom count, ligand-probe matching scores s_{pose}*,* s_{pose}^{*} *as defined in Eq. (78) and Eq. (79), and combinations of them using filtering thresholds;* x_o *and* x_{n4} *denote the matching with only o and n4 probe as defined in Eq. (77). The lines corresponding to the method yielding the best rankings (i.e. those where all three fragments are successfully retrieved at the beginning of the ranking) are marked in green. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/ fragments_screening/).*

	Fragment 1	Fragment 2	Fragment 3
ChemPLP	71	1	4
ChemPLP/N(HA)	1	41	32
s_{pose}	1	29	2
s_{pose}^{*}	1	6	2
ChemPLP ($s_{\mathrm{pose}} > 1/6$)	44	1	4
ChemPLP ($s_{\mathrm{pose}} > 1/4$)	24	1	4
ChemPLP ($s_{\mathrm{pose}} > 1/3$)	17	1	3
ChemPLP (x_{o} and $x_{\mathrm{n}} > 0.0$)	8	1	2
ChemPLP (x_{o} or $x_{\mathrm{n}} > 0.0$)	47	1	4
ChemPLP ($x_{\mathrm{n}} > 0.0$)	36	1	3
ChemPLP ($x_{\mathrm{o}} > 0.0$)	19	1	3
ChemPLP ($s_{\mathrm{pose}}^{*} > 0.0$)	8	1	2
ChemPLP/N(HA) ($s_{\mathrm{pose}} > 1/6$)	1	35	27
ChemPLP/N(HA) ($s_{\mathrm{pose}} > 1/4$)	1	28	23
ChemPLP/N(HA) ($s_{\mathrm{pose}} > 1/3$)	1	24	19
ChemPLP/N(HA) (x_{o} and $x_{\mathrm{n}} > 0.0$)	1	4	2
ChemPLP/N(HA) (x_{o} or $x_{\mathrm{n}} > 0.0$)	1	35	27
ChemPLP/N(HA) ($x_{\mathrm{n}} > 0.0$)	1	29	22
ChemPLP/N(HA) ($x_{\mathrm{o}} > 0.0$)	1	10	7
ChemPLP/N(HA) ($s_{\mathrm{pose}}^{*} > 0.0$)	1	4	2
s_{pose} (ChemPLP > 30)	1	26	2
s_{pose} (ChemPLP > 40)	1	17	2
s_{pose} (ChemPLP > 50)	-	4	1
s_{pose}^{*} (ChemPLP > 30)	1	6	2
s_{pose}^{*} (ChemPLP > 40)	1	3	2
s_{pose}^{*} (ChemPLP > 50)	-	2	1
s_{pose} (ChemPLP/N(HA) > 2)	1	29	2
s_{pose} (ChemPLP/N(HA) > 3)	1	23	2
s_{pose} (ChemPLP/N(HA) > 4)	1	-	-
s_{pose}^{*} (ChemPLP/N(HA) > 2)	1	6	2
s_{pose}^{*} (ChemPLP/N(HA) > 3)	1	4	2
s_{pose}^{*} (ChemPLP/N(HA) > 4)	1	-	-

Ranking w.r.t. the ligand-probe matching scores s_{pose} and s_{pose}^*, on the other hand, yields good retrieval for fragment 1 and 3 (ranks 1 and 2, respectively). Fragment 2 is scored on rank 29 for s_{pose} but is nicely retrieved by s_{pose}^* (rank 6). Thus, using only the restrictive s_{pose}^*, all the fragments can be found among the top ten scoring molecules. This underlines the benefit of the non-additive nature of the overlap score: it allows to retrieve fragments which match the thermodynamic binding site profile without being considerably biased towards larger or smaller molecules. Consequently, and opposed to the results for the virtual screening, usage of the ligand-probe matching score alone (without ChemPLP) here leads to more satisfactory results than usage of ChemPLP. As indicated, this results from the nature of the ChemPLP scoring function which was designed for classical virtual screening where molecules are usually not fragment-like.

To investigate the reasonability of the ligand-probe matching score, the poses of the other high scoring fragments ranked before fragment 2 according to s_{pose}^* were analysed; they are illustrated in Figure 39 in overlay with the ligand in 5c7a. The respective docking poses reveal that especially the fragment on rank 3 shows good overlap with the ligand in 5c7a w.r.t. functional groups: An ester carbonyl oxygen atom can be found in the same region where the amide carbonyl is located in 5c7a and which is highly favourable for accommodation of an oxygen substituent according to probe densities. Similarly, the nitrogen atoms in the five-ring are located in the area with high n4 probe density, albeit they are not charged, and the methyl group nicely overlaps with the methyl substituent of the 5c7a ligand.

The fragment on rank 4, too, shows nice overlay with the probe densities: It contains an amine located in the P1 pockets and bears a phosphate groups which, in the predicted pose, occupies an area of the binding site untargeted by the ligand in 5c7a which however exhibits high o probe densities (s. Figure 36).

The fragment on rank 5 contains a sulfonamide group in the respective region, as well as a carbonyl and a methyl group that overlay with their counterparts in the 5c7a ligand. Thus, the respective fragments can be considered reasonable “hits” w.r.t. matching the thermodynamic binding site profile. They even show the versatility of the concept since they are examples of structurally completely dissimilar molecules of different size which nevertheless show good matching with the probe densities.

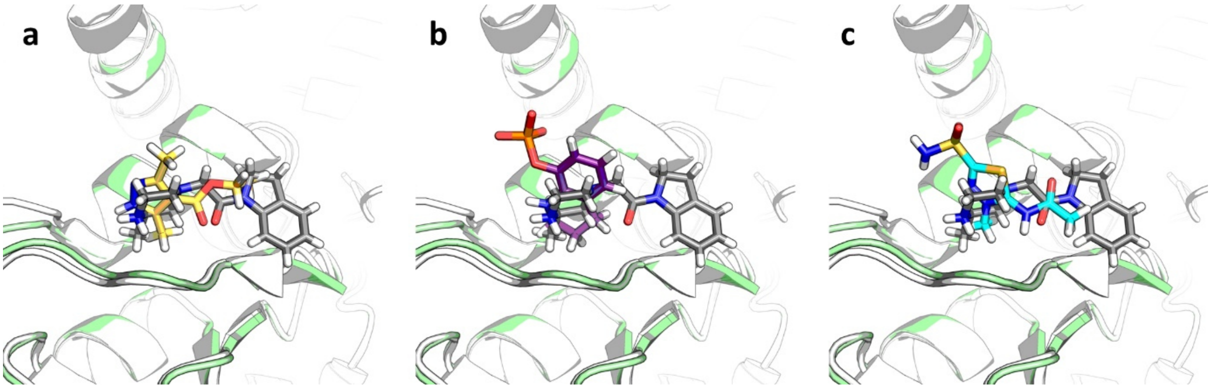

Figure 39: Docking poses of fragments with a high ligand-probe matching score in 3hl5 (pale green) in overlay with the complex structure 5c7a (protein in white, ligand in grey); a) fragment 61 (rank 2 by s_{pose}^*), b) fragment 35 (rank 4 by s_{pose}^*), c) fragment 47 (rank 5 by s_{pose}^*). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/ fragments_screening/ (fragment 61: pose 61_1; fragment 35: pose 35_16; fragment 47: pose 47_25)).

Despite the good performance of s_{pose}^*alone, it was investigated if - like for the virtual screening - an improved retrieval of the three fragments of interest can be achieved by combination of ChemPLP and the ligand-probe matching score via filtering. The results (Table 25) show that ranking of molecules by ChemPLP after filtering based on absolute s_{pose} values results in a better rank for fragment 1 (from 71 to 44, 24, and 17 for threshold 1/6, 1/4, and 1/3) which is however still not satisfactory. Only when demanding that x_{o} and $x_{\text{n}} > 0.0$ or $s_{\text{pose}}^* > 0.0$, fragment 1 can be found among the top 10 ranked molecules. This indicates that, when dealing with fragment-like molecules, a stricter filtering criterion w.r.t. the ligand-probe matching score has to be applied than in the classical virtual screening scenario.

Similar results are obtained for ranking according to ChemPLP normalised by the heavy atom count after filtering based on overlap score: Filtering by absolute s_{pose} values > 1/3 improves the ranks of fragment 2 and 3 from 41 and 32 to up to 24 and 19. Ranking among the top 10 molecules is achieved when demanding that $x_{\text{o}} > 0.0$ or that x_{o} and $x_{\text{n4}} > 0.0$, with the latter yielding almost perfect retrieval of the three fragments within the top four ranked molecules. The same result is achieved via combination of normalised ChemPLP and filtering by s_{pose}^*, which thus slightly outperforms ranking by ligand-probe matching score alone. The combination with s_{pose}^* already lead to almost perfect active-decoy differentiation in virtual screening, underlining that a strict filtering according to probe density followed by (potentially normalised) conventional scoring is a promising way of post-processing docking results.

Filtering molecules according to conventional ChemPLP score, followed by ranking by s_{pose}, leads to moderate improvement for thresholds of 30 and 40 (ChemPLP). For higher thresholds, however, fragment 1 is sorted out due to its small size and thus small absolute ChemPLP score. A similar trend is observed for filtering according to the normalised ChemPLP score: here, fragments 2 and 3 are sorted out for high thresholds due to their higher number of heavy atoms. When combining the respective filtering with s_{pose}^{*}-based ranking, ideal or nearly ideal retrieval is achieved for ChemPLP > 40 and ChemPLP/N(HA) > 3.

In summary, the results for the fragment library screening show that the ligand-probe matching score is a suitable means to retrieve fragments of any size or chemical structure which match the thermodynamic binding site profile. For optimal ranking power, it should be employed as a filtering criterion together with classical scoring. In contrast to classical virtual screening, the filtering criterion should be rather strict for fragment-like molecules to compensate the dependence of classical scoring functions on the molecule size. Thus, new chemical scaffolds can be found as starting points for the *de novo* design of novel ligands.

4.3 Application of RISM-based descriptors for SBDD – case studies

In the previous chapters, RISM-based approaches for the local characterisation of binding sites and protein-ligand interactions were introduced, and proof-of-concept studies as well as validation studies on available data sets were presented. In this chapter, the developed methods will be applied to specific case studies to illustrate their benefit on existing challenges relevant in medicinal chemistry.

A particularly important yet highly demanding field in this context are protein-protein interactions (PPI). They play an important role in all physiological processes; hence, their dysregulation is involved in many diseases. Yet, there are so far only few therapeutic agents on the market that target PPIs since their characteristics make them challenging targets for drug design: PPI interfaces are usually flat and large, and no natural small molecule binding partners are available that could serve as a starting point for rational drug design.[304]

In this chapter, the utility of the RISM-based water analysis and pharmacophoric probe densities will be investigated w.r.t. the characterisation of PPI interfaces and the design of respective binding partners on the basis of three exemplary proteins that are relevant to medicinal chemistry: In a first case study, both the local water and probe thermodynamics will be used to retrospectively evaluate the development of a hit-to-lead series of inhibitors of the protein XIAP which is an important regulator of apoptosis. In a second example, the respective approaches are applied to develop a design strategy for the generation of Ugi-type inhibitors of Bcl-xL, a highly challenging target that exhibits considerable structural flexibility. In a third case study, the binding site of the protein hTEAD, a regulator of the Hippo pathway relevant for cell proliferation, is characterised to explain SAR trends on modified peptide binding partners and to prospectively guide the design of respective screening libraries: Thanks to an advanced technology developed there, the Brunschweiger group has a broad portfolio of chemical reactions that can be applied for the synthesis of large, combinatorial DNA-encoded libraries.[305] In this work, an attempt was made to support the selection of especially promising starting building blocks for these libraries.

All raw data for the analyses in 4.3 can be found in the Electronic Appendix (folders XIAP/, Bcl-xL/, and TEAD/). This includes respective ligand and protein structures (ligand.pdb, pocket.pdb), the calculated water positions with thermodynamic properties (Ghyd@water_apo.pdb, Ghyd@water_holo.pdb), as well as the interpolated *apo* water thermodynamic data on the ligands (Ghyd@lig.pdb) and the interpolated probe *g*-function data on the ligands (gUU_xy@lig.pdb). For examples for which probe densities are shown in this work, also respective cube files are provided.

4.3.1 XIAP – Analysing a hit-to-lead series

The benefit of RISM-derived thermodynamic binding sites profiles for ligand optimisation is best illustrated for a set of closely related compounds, ideally in a hit-to-lead series. Analysis of the *holo* binding site water molecules for different ligands can for instance reveal conserved hydration sites but also newly introduced or destabilised sites. The respective ligand parts can then be modified, e.g. by introducing larger groups to replace an isolated, unstable water position. Likewise, the pharmacophoric probe densities in the direct proximity of a hit molecule show where respective substituents could be added, thus also providing valuable information w.r.t. to design directions.

A suitable example that allows to investigate the utility of both the water analysis and the pharmacophoric probe densities is the protein XIAP, which was already used in the proof-of-concept studies in 4.2. The availability of several respective complex structures, including a hit-to-lead series,[261] and the high relevance of this protein for the treatment of cancer make it an ideal model system for this work. In the following, both the water thermodynamics and the probe densities will be discussed for the respective structures w.r.t. their usefulness for explaining the corresponding SAR trends and design strategy. Throughout the whole analysis, the same algorithms and settings were used as for the data sets in 4.1 and 4.2, i.e. hydration site positions and their $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions were calculated as described in 3.4, and the probe densities were calculated and analysed as outlined in 3.2 and 3.5.

The starting point for the study is the complex structure 5c3h with a hit compound (IC_{50} > 5000 µM) containing a piperidine ring and a piperazine ring. The latter binds in the same region as the N-terminal Ala residue of XIAP's natural peptide binding partners.[303] In Figure 40, the respective complex structure is shown together with the predicted *apo* water molecules and the interpolated c3, n4, and o probe densities, similar to Figure 36 in 4.2.1.

It can be seen that the P1 pocket, which accommodates the piperazine ring, contains water molecules with highly favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions. This is in line with the finding from 4.1 that ligand atoms of the n4 type are found to replace almost exclusively "happy" water molecules. Consequently, the position of the respective amine in the piperazine ring coincides with an n4 probe maximum. Since this is the position where the terminal amino group of the Ala residue in XIAP's natural binding partners is usually accommodated, this is highly intuitive. As already shown for 5c7a in 4.2, there is also a nice match between o probe density and the carbonyl group in the ligand in 5c3h. Its position in the P2 pocket corresponds to the location of the backbone carbonyl in XIAP's peptide binding partners and can undergo an H-bond with a neighbouring Thr residue.[303] High interpolated c3 probe values can be observed for the piperidine ring which is located in the P3 pocket where usually a Pro sidechain is accommodated.[303] Thus, the probe and water-based analysis reveals a good match between the ligand structure and the binding site profile, which is reasonable since the ligand mimics several characteristics of the natural binding partner peptides.

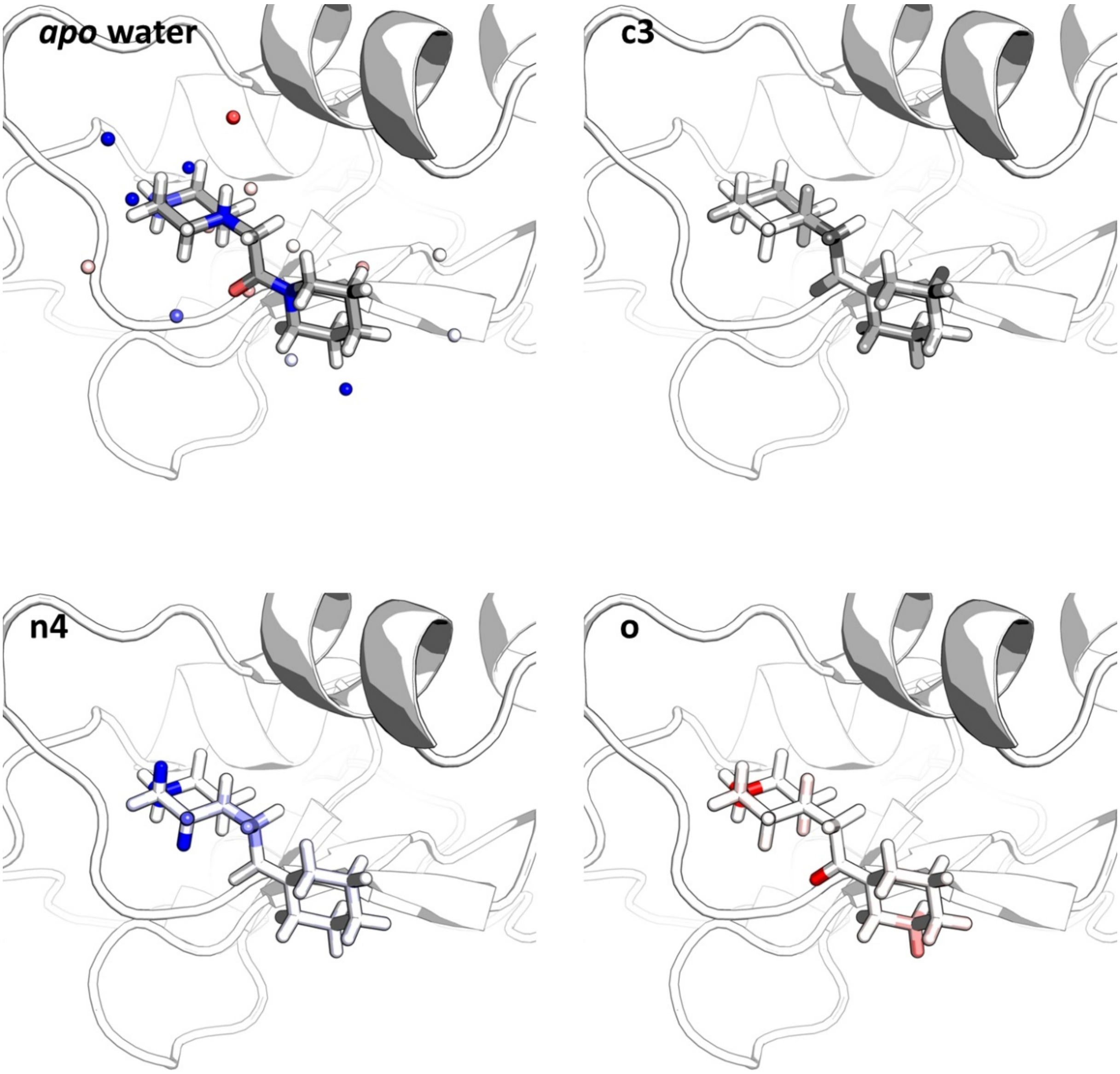

Figure 40: Apo water thermodynamics and interpolated apo c3, n4, and o probe g-function values on XIAP complex structure 5c3h. Predicted apo water molecules are coloured by their calculated $\Delta_{hyd}G_P$ contributions (from blue to red from -2.0 to +2.0 in units of kcal/mol); c3 probe: colouring from white to grey from 0 to 2, n4 probe: colouring from white to blue from 0 to 30; o probe: colouring from white to red from 0 to 10). Water molecules within 3.5 Å of the ligand are shown. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/ 5c3h/).

To get an idea how the ligand in 5c3h can be further optimised, the respective water thermodynamics and probe densities of the *holo* complex can be investigated. In Figure 41, the predicted *holo* water thermodynamics in the proximity of the piperazine ring are shown.

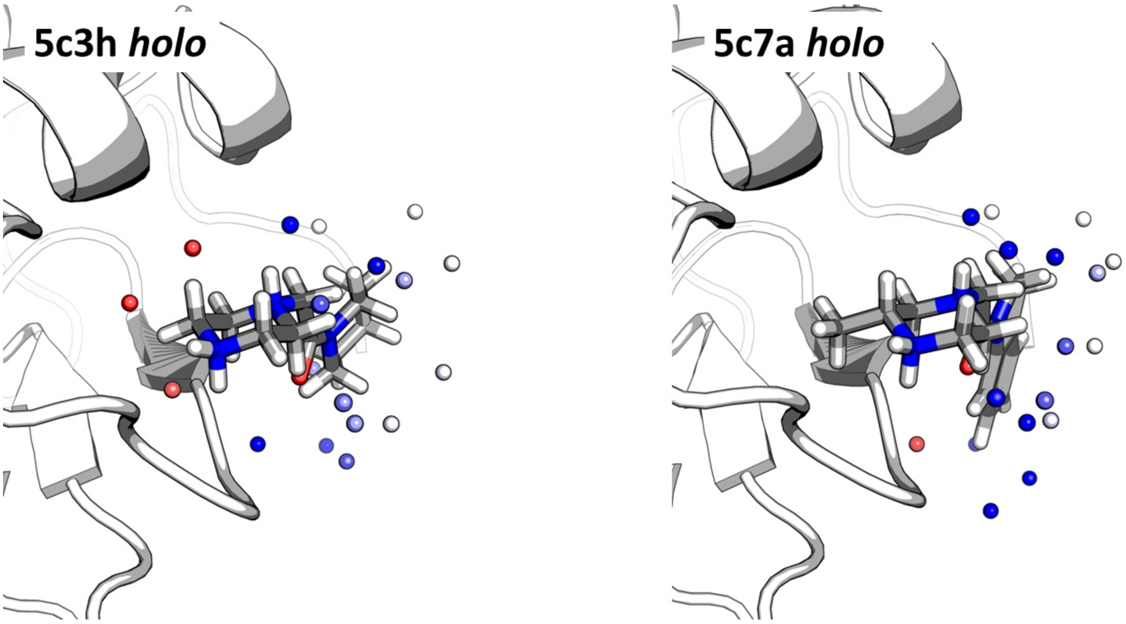

Figure 41: Holo water thermodynamics in XIAP complexes 5c3h and 5c7a. Predicted holo water molecules are coloured by their calculated $\Delta_{hyd}G_P$ contributions (from blue to red from -2.0 to +2.0 in units of kcal/mol). Water molecules within 3.5 Å of the ligand are shown. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/) in the respective pdb folders.

Intriguingly, although the respective amine is ideally accommodated by the P1 pocket, 3D RISM-based algorithms predict the presence of several highly unstable hydration sites in the proximity of the piperazine ring. Hence, replacement of these rather isolated, unstable hydration sites via introduction of suitable substituents should be favourable.

Fortunately, a closely related compound with an additional methyl group at the piperazine ring was synthesised and co-crystallised by Chessari *et al*., leading to complex structure 5c7a (Figure 41).[261] Indeed, in this *holo* structure, two destabilised water positions are eliminated. The exact influence of this water replacement on affinity cannot be determined since, in addition to introduction of the methyl group, the piperidine ring in the hit compound was changed to an indole ring. However, the large overall gain in affinity (IC_{50} > 495 µM) can be seen as a hint that replacement of the newly introduced unstable hydration sites has a favourable effect.

In Figure 42, the corresponding c3 probe densities for 5c3h are depicted in comparison with the respective modified ligand 5c7a. Indeed, there is a large area with *holo* c3 probe density next to the piperidine ring, suggesting there is space to accommodate larger apolar groups in this region. Consequently, the 5c7a ligand, which contains an indole ring instead of the piperidine ring, nicely overlaps with this density. Interestingly, at the chosen thresholds, no additional peak could be observed in the area where the methyl group is added in 5c7a and which showed clear optimisation potential w.r.t. water thermodynamics. At the same time, no significant high energy water molecules were predicted in the region around the piperidine ring (Figure 41). This nicely demonstrates that both approaches, the water and probe analysis, can complement each other and together provide an in-depth thermodynamic picture that can be used for rational ligand optimisation and SAR studies.

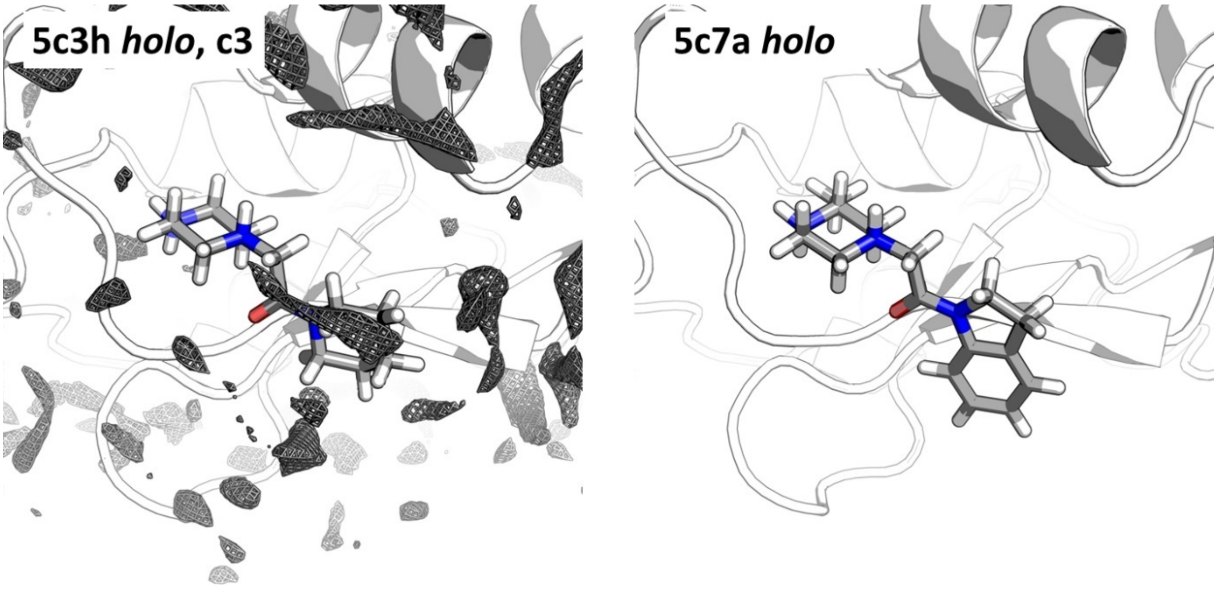

Figure 42: Complex 5c3h with holo c3 probe densities (threshold: 10) and complex 5c7a of the modified ligand. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/ 5c3h/).

In a next step, the ligand of 5c7a was modified by Chessari *et al*. via addition of a chlorine substituent at the 6-position and two methyl groups at the 3-position of the indole moiety, which leads to a hundredfold improvement w.r.t. inhibition (IC_{50} = 5.5 µM). To see if these or similar modifications could have been proposed by the water and probe approaches, the respective *holo* results for 5c7a can be evaluated: The water analysis (Figure 43) predicts the presence of a newly introduced unstable water molecule near the six-ring of the indole moiety and thus indeed suggests the addition of a large substituent at the 6-position. Based on the analysis in 4.1, a halogen atom or a methyl group could be considered a suitable group, which thus is in nice agreement with the found SAR. Consequently, the respective analysis on the complex with the modified compound (5c7c, right panel in Figure 43) indeed reveals that the unstable water molecule is eliminated in the presence of the chlorine substituent. Comparison with the respective *apo* binding site situation shows that the water molecules in the area of the chlorine substituent exhibit slightly positive $\Delta_{\text{hyd}}G_{\text{P}}$ contributions also in the *apo* form. Thus, the overall replacement here is roughly in line with the rules derived in 4.1.1, although the trend is not as pronounced as for the MMP examples in 4.1.3. The hydration sites near the 5-ring in 5c7a, which are replaced or shifted due to the presence of the methyl groups in 5c7c, show moderate $\Delta_{\text{hyd}}G_{\text{P}}$ contributions, suggesting that they are not particularly unstable but should be rather easy to replace.

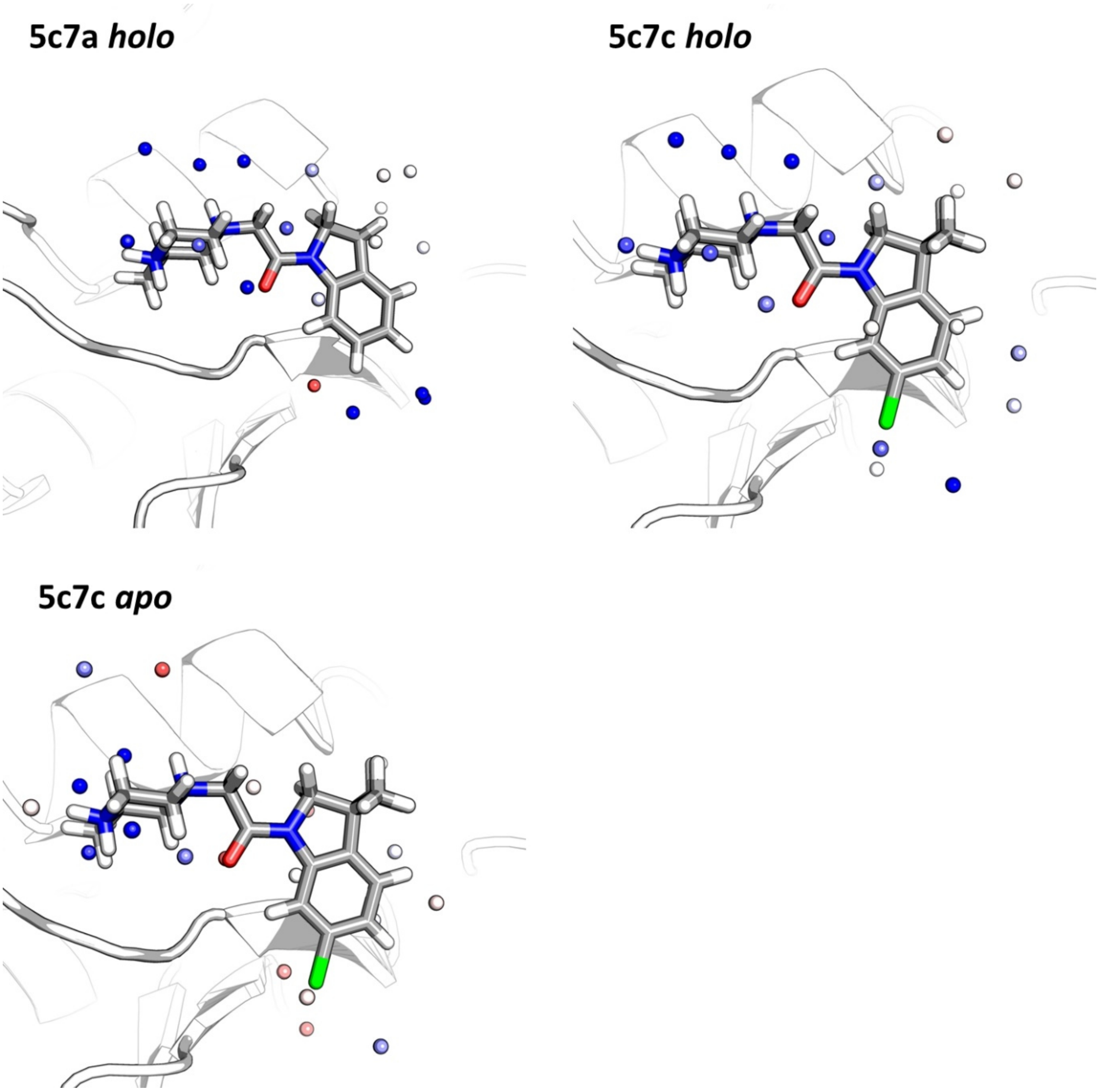

Figure 43: Holo water thermodynamics in XIAP complexes 5c7a and 5c7c as well as apo water thermodynamics of 5c7c for comparison. Predicted holo water molecules are coloured by their calculated $\Delta_{hyd}G_{P}$ contributions (from blue to red from -2.0 to +2.0 in units of kcal/mol). The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/) in the respective pdb folders.

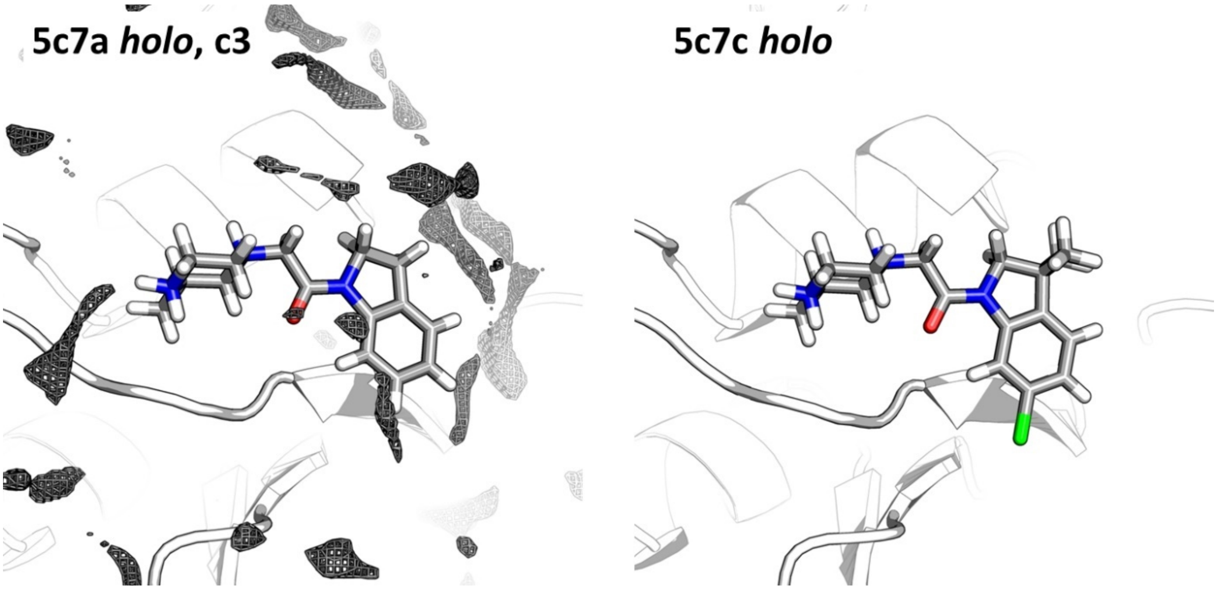

Figure 44: Complexes 5c7a with holo c3 probe densities (threshold: 10) and 5c7c of the modified ligand. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/) in the respective pdb folders.

The *holo* c3 probe density of 5c7a (Figure 44) shows peaks around the indole moiety, in the area where the newly introduced methyl groups and the chlorine substituent are located. Thus, the probe density is again in good agreement with the design strategy by Chessari *et al*.

From 5c7c to 5c83 (IC_{50} = 0.16 μM) and to 5m6h (IC_{50} = 0.15 μM),[306] rather large modifications are introduced (addition of a phenyl group and an aliphatic ether group/morpholine amide. In 5c7c, (Figure 45) there are respective c3 probe peaks above the indole ring and in the proximity of the chlorine substituent, which nicely overlap with the newly introduced phenyl ring and the terminal methyl group of the ether substituent.

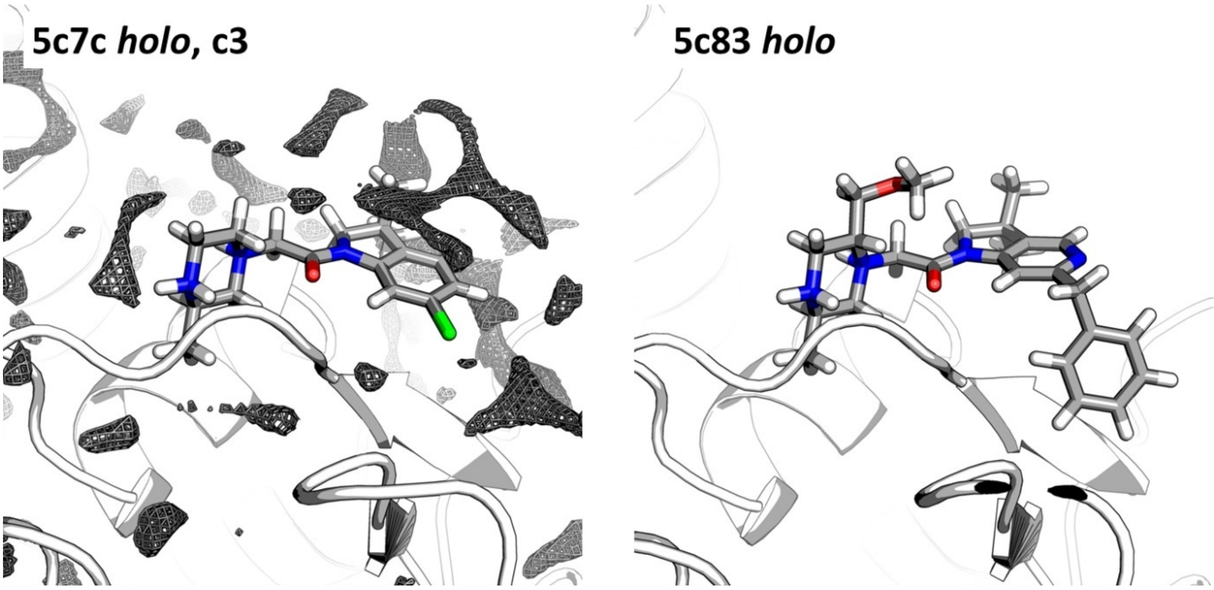

Figure 45: Complexes 5c7c with holo c3 probe densities (threshold: 10) and 5c83 of the modified ligand. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/) in the respective pdb folders.

The ligand in 5m6h is not more potent than the one complexed in 5c83 but bears an alternative morpholine moiety in the region of the methoxy group. In Figure 46, the *holo* c3 and o probe density in 5c83 is shown. The c3 probe density is mainly located around the methyl group, while the respective o probe density nicely overlaps with the carbonyl oxygen atom position of the modified substituent. The o peak in this area is not obvious at first glance since there are no binding site residues in the direct proximity, and the morpholine moiety in the respective ligand rather points out of the binding site. However, analysis of the respective crystal structure 5m6h and the predicted *holo* water thermodynamics (Figure 47) reveals the presence of a water molecule near the carbonyl group that has a highly favourable $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contribution and bridges hydrogen bonds between the oxygen atom and a near Gln and Trp residue.

The presented analysis thus shows that the RISM-based water analysis and pharmacophoric probe densities can be employed for guiding the further optimisation of hit compounds. While the c3 probe density peaks rather reveal overall regions where larger, apolar groups can be added, the o and n4 probe peaks hint at areas that can be exploited for hydrogen bonds or ionic interactions. As the above presented example shows, it also reveals less obvious regions where respective groups can undergo bridged hydrogen bonds via neighbouring water molecules. The respective RISM-based water analysis can be employed to identify such bridging water molecules but also to reveal specific, often newly introduced unstable hydration sites which can then be specifically targeted for replacement by tailored modifications.

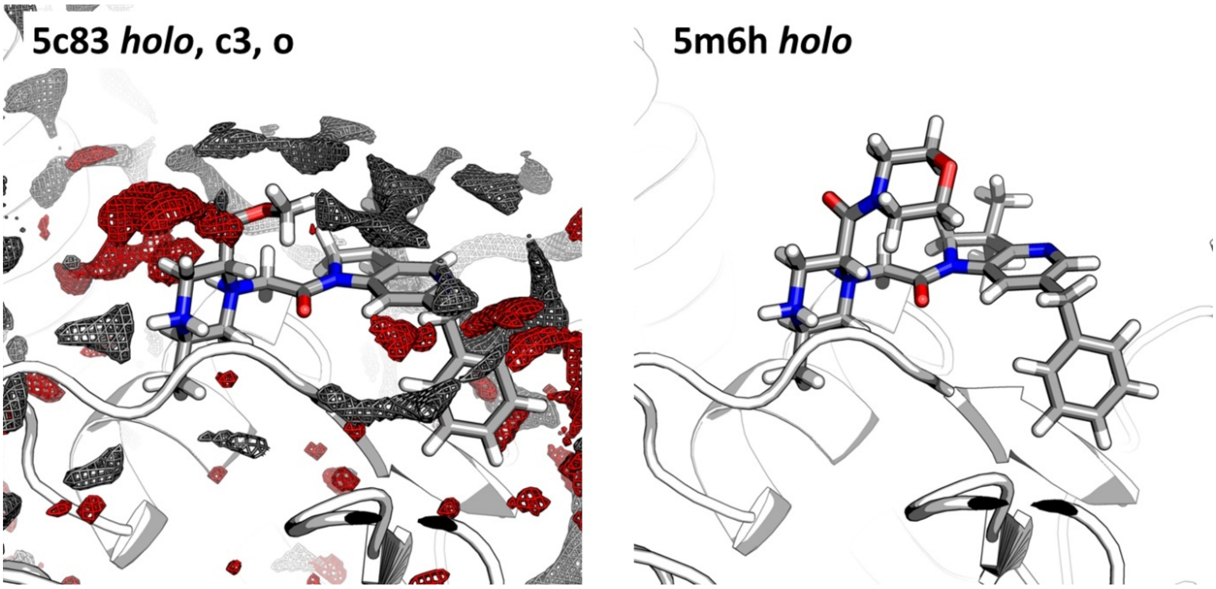

Figure 46: Complex 5c83 with holo c3 (grey, threshold: 10) and o (red, threshold: 400) probe densities and complex 5m6h of the modified ligand. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/) in the respective pdb folders.

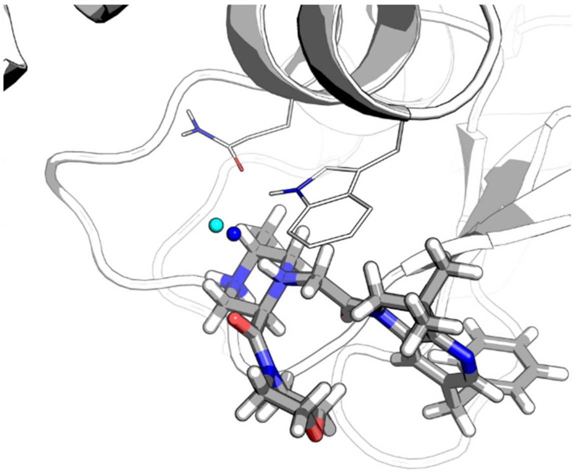

Figure 47: Experimental (cyan) and predicted holo water position (colouring according to the calculated $\Delta_{hyd}G_P$ contributions from blue to red from -2.0 to +2.0 in units of kcal/mol) in 5m6h that bridges an interaction between the carbonyl group and an Trp and Gln residue of XIAP. The respective raw data can be found in the Electronic Appendix (Electronic Appendix/ XIAP/5m6h)

4.3.2 Bcl-xL – Deriving design strategies for DNA-encoded libraries

As a regulator of apoptosis, Bcl-xL and other proteins of the Bcl2 family are attractive targets for the treatment of cancer. Abbot Laboratories developed orally available inhibitors of the Bcl-2 family based on an N-acylsulfonamide scaffold.[307] Dömling *et al*. could show that it is possible to replace this N-acylsulfonamide scaffold by the α-acylaminocarboxamide backbone of the Ugi reaction (Figure 48, Figure 49), which is highly attractive due to the relatively fast and easy synthesis and the circumvention of issues concerning intellectual property.[308] Furthermore, the Ugi reaction is an attractive synthesis route for combination with DNA-encoded libraries as designed by Kunig *et al*.[309]

However, the respective compounds exhibit binding in the low micromolar range and thus are at least 2 to 3 orders of magnitude less potent than the Abbot compounds.[308] Hence, it is highly attractive to optimise the choice and design of the Ugi reaction components to improve the interactions of the α-acylaminocarboxamide-based compounds in the binding site of Bcl-xL.

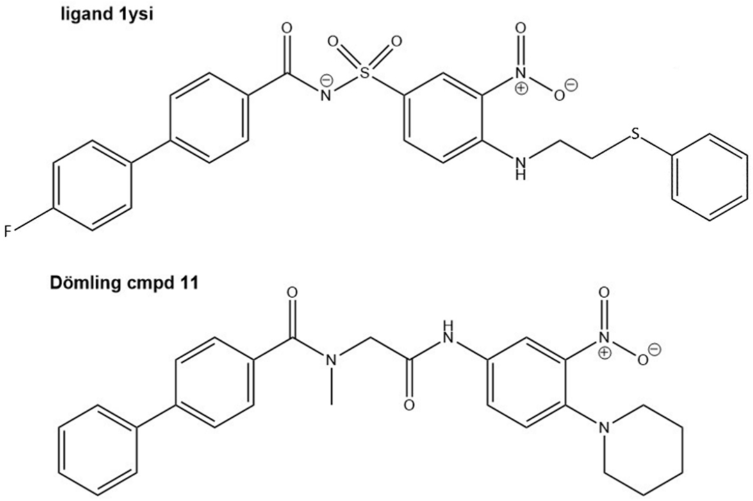

Figure 48: Exemplary Bcl-xL N-acylsulfonamide-based inhibitor by Abbot Laboratories (ligand 1ysi)[307] and a representative Ugi-based compound synthesised by Dömling et al[308] (Dömling cmpd 11).

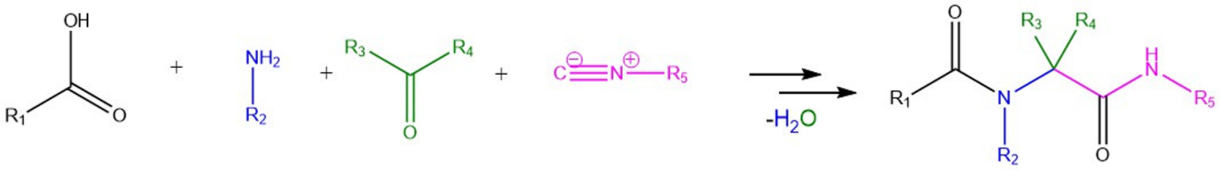

Figure 49: Schematic illustration of the Ugi multicomponent reaction.

The aim of the work presented in this chapter therefore was to generate respective design strategies for a Bcl-xL-tailored DNA-encoded library on the basis of two Ugi model compounds (Figure 50) using RISM-based binding site characterisation. The studies shown here are based on analyses and MD simulations performed by J. Borchert in the Kast group under guidance of the author.[310] Structures that were generated within the scope of these studies were used here to perform the RISM-based analyses introduced in 3.5 and 3.6 to get insights w.r.t. further ligand design.

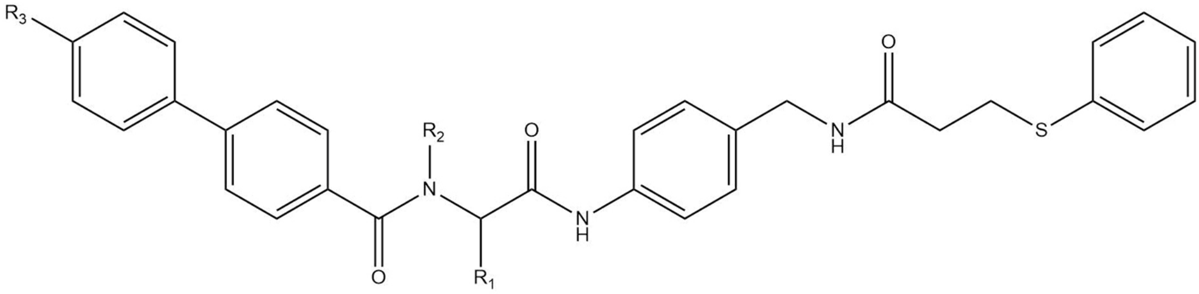

Figure 50: Chemical structure of the Ugi model compounds; compound 1: R1 = H, R2 = Me, R3 = F; compound 2: R1 = Et, R2 = n-Pr, R3 = H.

Derived from the Abbot inhibitors, the two model compounds contain a biphenyl group and a phenylsulfanyl group; the latter can be introduced via an amide coupling after the Ugi reaction. The phenyl ring corresponds to the nitro-phenyl group in the Abbot inhibitors, and the N-acylsulfonamide-based backbone is replaced by the α-acylaminocarboxamide backbone of the Ugi reaction using different substituents for the R1 and R2 position.

Bcl-xL is a challenging target with a highly flexible binding site that undergoes large structural changes upon ligand binding: Overlay of representative structures (Figure 51) shows that the protein adapts depending on the bound ligand, so that for instance the ligand as bound in complex structure 4c5d[311] would collide with the protein's conformation in the complex structure 1ysi.[307] This high flexibility makes the structural design of ligands highly challenging since it hampers the use of classical tools like docking with a rigid protein binding site.

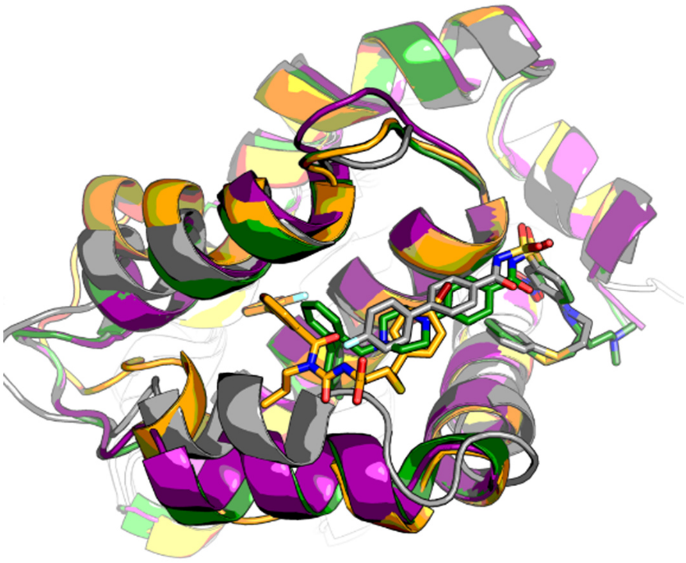

Figure 51: Flexibility of the Bcl-xL binding site: overlaid structures 1ysi[307] (grey), 2yxj[312] (green), 4c5d[311] (orange), and 4cin[313] (purple) (peptide bound in 4cin not shown).

Therefore, a prerequisite for the RISM-based binding site characterisation w.r.t. Ugi-based compounds is the availability of suitable protein conformations. Until today, no crystal structure has been solved of Bcl-xL with an Ugi-based inhibitor. Therefore, the needed binding site conformations especially tailored towards Ugi-based compounds were generated within the scope of a Bachelor thesis by J. Borchert under guidance of the author using a multistep workflow comprising MD simulation and docking (Figure 52). In this workflow, an MD simulation of the complex structure 1ysi with an N-acylsulfonamide-based inhibitor (Figure 48) from which the respective Ugi model compounds were derived was performed first. The respective trajectory was clustered, and a representative structure of the highest populated cluster was used for docking of the two Ugi model compounds (Figure 50) into Bcl-xL. The top docking poses of both Ugi model compounds were then again submitted to MD simulation, and a representative structure of the highest populated cluster was determined.[310]

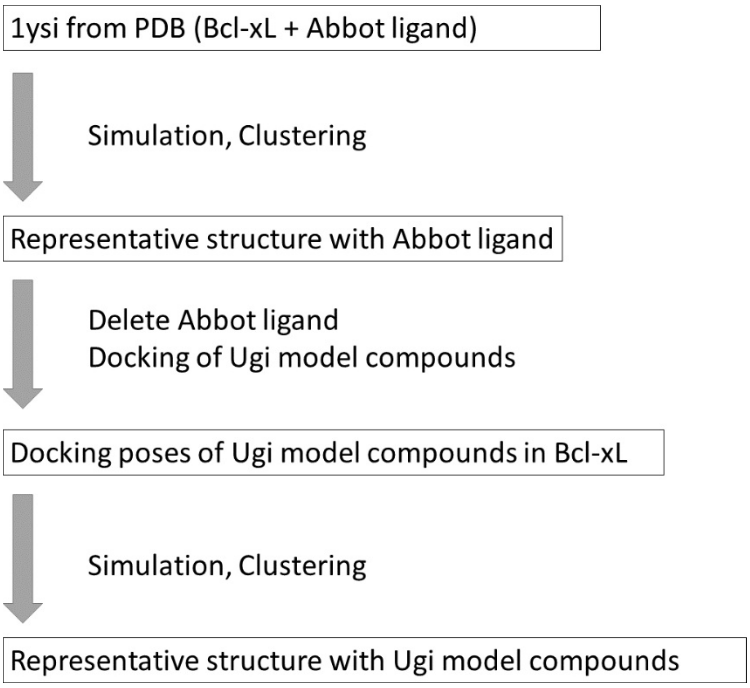
1ysi from PDB (Bcl-xL + Abbot ligand)
Simulation, Clustering
Representative structure with Abbot ligand
Delete Abbot ligand
Docking of Ugi model compounds
Docking poses of Ugi model compounds in Bcl-xL
Simulation, Clustering
Representative structure with Ugi model compounds

Figure 52: Overview of the workflow that was employed by J. Borchert to obtain complex structures of the Ugi model compounds in Bcl-xL.

In this work, the representative structures of Bcl-xL with two Ugi-type model compounds obtained by J. Borchert's studies were used for RISM-based binding site characterisation with the aim to draw design conclusions w.r.t. the building blocks for the different components of the Ugi reaction, i.e. the carboxylic acid, amine, aldehyde, and isonitrile part (s. Figure 49), as well as the sulfanylphenyl part which is added after the Ugi reaction via amide coupling. For instance, it is important to characterise where the respective substituents are located within the binding site and what kind of interactions they can undergo. In addition, for use of the Ugi-based compounds in DNA-encoded libraries, an anchor position for a linker to the DNA barcode as unique identifier of the compounds it needed. The positioning of this linker in the molecule is of utmost importance since addition at a region which is deeply buried in the binding site or undergoes vital interactions with a protein residue would impede binding and would thus generate false negatives during the experimental screening. Therefore, finding the optimal position where a linker can be introduced in all library compounds was also part of the studies carried out by J. Borchert and the further analysis in this work.

In Figure 53, the input structures for the MDs (white) carried out by J. Borchert and the representative structure of the highest populated cluster of the resulting MDs (pale green) are shown for a) the Abbot inhibitor from 1ysi, b) the Ugi model compound 1 docked into the representative structure of the Abbot MD, and c) the Ugi model compound 2 docked into the representative structure of the Abbot MD (s. 3.6). It has to be noted that, due to introduction of the R1 substituent, model compound 2 has a chirality centre, and both enantiomers can be obtained by the Ugi reaction. In J. Borchert's work, all simulations were carried out with the R-enantiomer for which the ethyl group nicely points out of the binding pocket in the starting conformation. Therefore, the analyses presented here are also limited to the R-enantiomer for this compound.

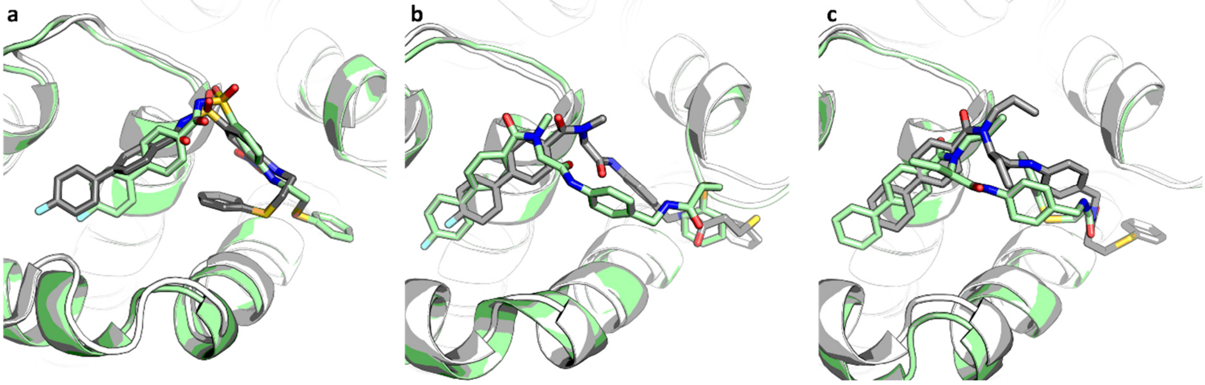

Figure 53: Overlay of MD starting structures (white) and the representative structure of the highest populated cluster of the MD for a) the Abbot inhibitor in 1ysi; b) Ugi model compound 1 in the representative structure of the Abbot inhibitor MD; c) Ugi model compound 2 in the representative structure of the Abbot inhibitor MD. The respective raw data can be found in the electronic appendix (Electronic_Appendix/Bcl-xL/) in the respective structure folders (1ysi, rep_1ysi, rep_ugi, rep_ugi_2).

Comparison of the structures shows that during the MD simulation, all three ligands undergo rather large conformational changes. In case of the Abbot ligand, the phenylsulfanyl group switches from a stacked conformation to an extended conformation, implying that this region is rather flexible. The docking poses of both Ugi ligands in the representative structure of the Abbot MD exhibit a similar, elongated conformation, while in the representative cluster structures, the phenylsulfanyl group assumes a different conformation. In addition, also the α-acylaminocarboxamide backbone and the neighbouring ring system show rather large structural changes compared to the starting structure, emphasising that the Bcl-xL binding site is difficult to tackle when neglecting flexibility.

To examine how the presence of the different ligands and their different conformations influence the overall thermodynamic binding site profile, RISM-based water thermodynamics and pharmacophoric probe densities were calculated for the MD starting structures and representative cluster structures for all three ligands. The resulting *apo* water molecules and interpolated probe densities are given in Figure 54 to Figure 56.

In the following, the respective results will be discussed to draw conclusions for the optimal choice of the respective building blocks, i.e. the isonitrile, carboxylic acid, aldehyde, and amine part as well as the amide coupling part containing the phenylsulfanyl moiety.

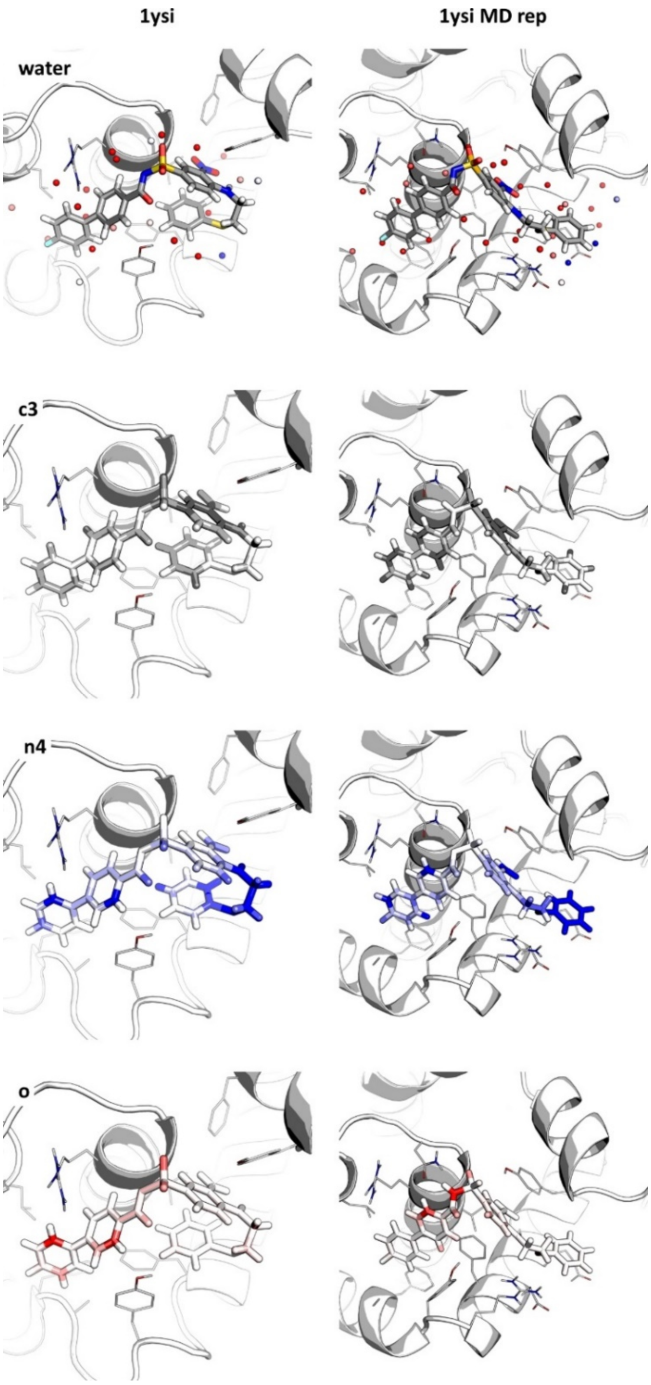

Figure 54: Apo water thermodynamics and interpolated apo c3, n4, and o probe g-function values for Bcl-xL structure 1ysi as taken from the PDB (left) and for the respective cluster representative of the MD simulation with the native ligand (right). Water molecules within 3.5 Å of the ligand are shown and coloured by their calculated $\Delta_{hyd}G_P$ contributions (from blue to red from -2.0 to +2.0 in units of kcal/mol); probe colouring: c3 - from white to grey from 0 to 3; n4 - from white to blue from 0 to 50; o - from white to red from 0 to 0.3. The different absolute ranges result from the charge of the protein (-12), leading to higher densities for the positively charged n4 probe and lower ones for the o probe. The respective raw data can be found in the electronic appendix (Electronic_Appendix/Bcl-xL/) in the respective structure folders (1ysi, rep_1ysi).

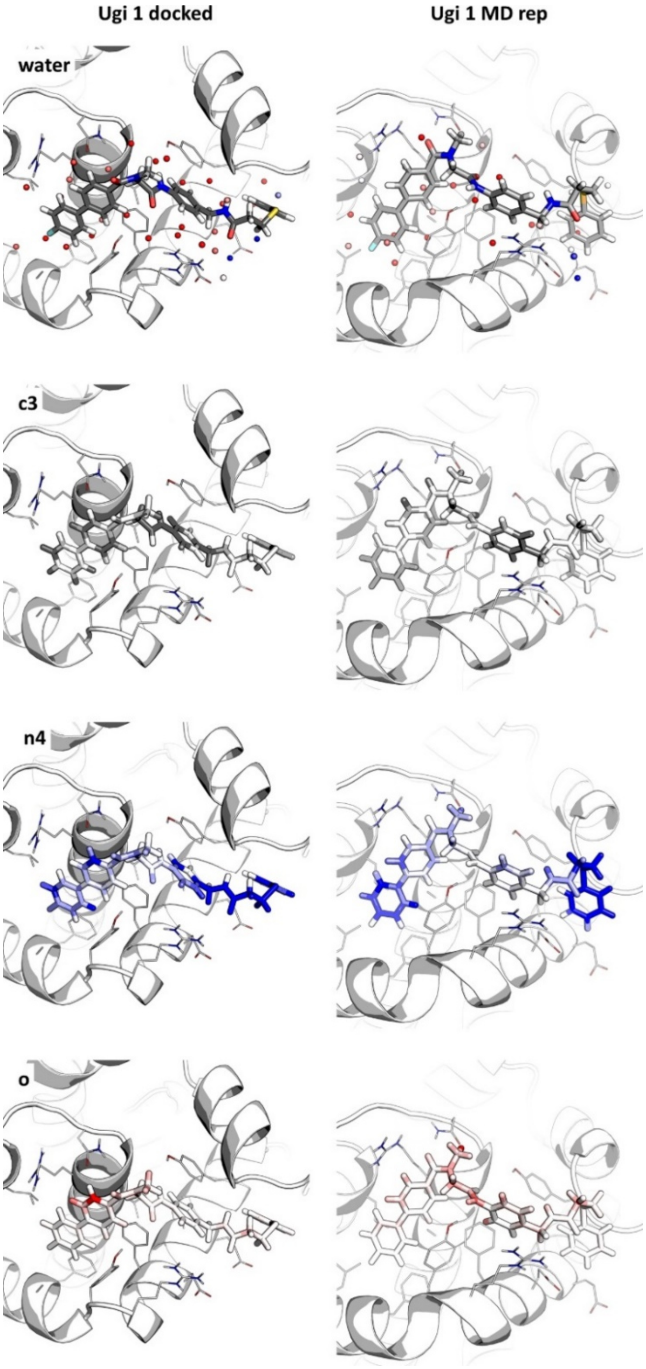

Figure 55: Apo water thermodynamics and interpolated apo c3, n4, and o probe g-function values for Ugi model compound 1 as docked into the representative structure of 1ysi (left) and for the respective cluster representative of the MD simulation (right). Water molecules within 3.5 Å of the ligand are shown and coloured by their calculated $\Delta_{hyd}G_P$ contributions (from blue to red from -2.0 to +2.0 in units of kcal/mol); probe colouring: c3 - from white to grey from 0 to 3; n4 - from white to blue from 0 to 50; o - from white to red from 0 to 0.3. The different absolute ranges result from the charge of the protein (-12), leading to higher densities for the positively charged n4 probe and lower ones for the o probe. The respective raw data can be found in the electronic appendix (Electronic_Appendix/Bcl-xL/) in the respective structure folders (rep_1ysi, rep_ugi_1).

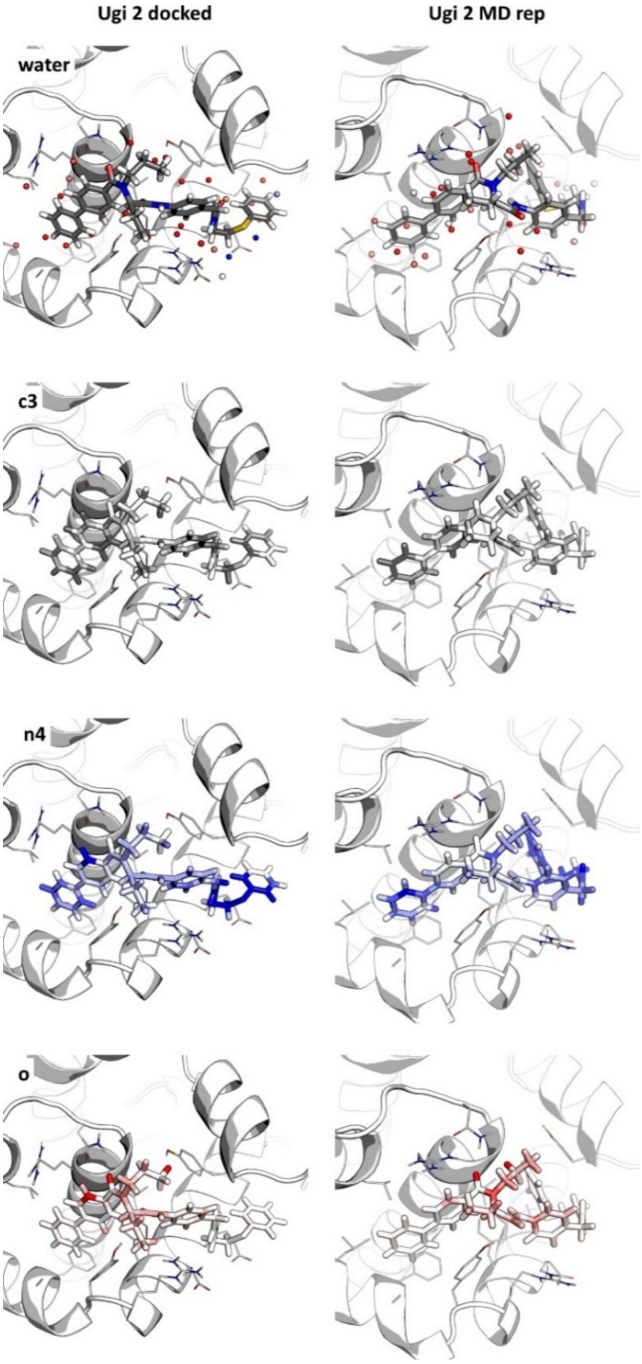

Figure 56: Apo water thermodynamics and interpolated apo c3, n4, and o g-function values for Ugi model compound 2 as docked into the representative structure of 1ysi (left) and for the respective cluster representative of the MD simulation (right). Water molecules within 3.5 Å of the ligand are shown and coloured by their calculated $\Delta_{hyd}G_P$ contributions (from blue to red from -2.0 to +2.0 in units of kcal/mol); probe colouring: c3 - from white to grey from 0 to 3; n4 - from white to blue from 0 to 50; o - from white to red from 0 to 0.3. The different absolute ranges result from the charge of the protein (-12), leading to higher densities for the positively charged n4 probe and lower ones for the o probe. The respective raw data can be found in the electronic appendix (Electronic_Appendix/Bcl-xL/) in the respective structure folders (rep_1ysi, rep_ugi_2).

4.3.2.1 Carboxylic acid part

Inspired by the Abbot inhibitors, the carboxylic acid building block introduces a biphenyl moiety into the Ugi model compounds. While the eastern part of all investigated ligands (based on the orientation in Figure 50 and Figure 51), especially the region around the phenylsulfanyl moiety, shows large structural changes from MD starting structure to the representative cluster structure (Figure 53 to Figure 56), the biphenyl ring seems to be a stable anchor. As can be seen in Figure 57 for the representative structure of the simulations with the Abbot inhibitor, and as implied by the throughout high interpolated c3 probe values in all structures (Figure 54 to Figure 56), it is located in a large, hydrophobic pocket with multiple high energy *apo* water sites.

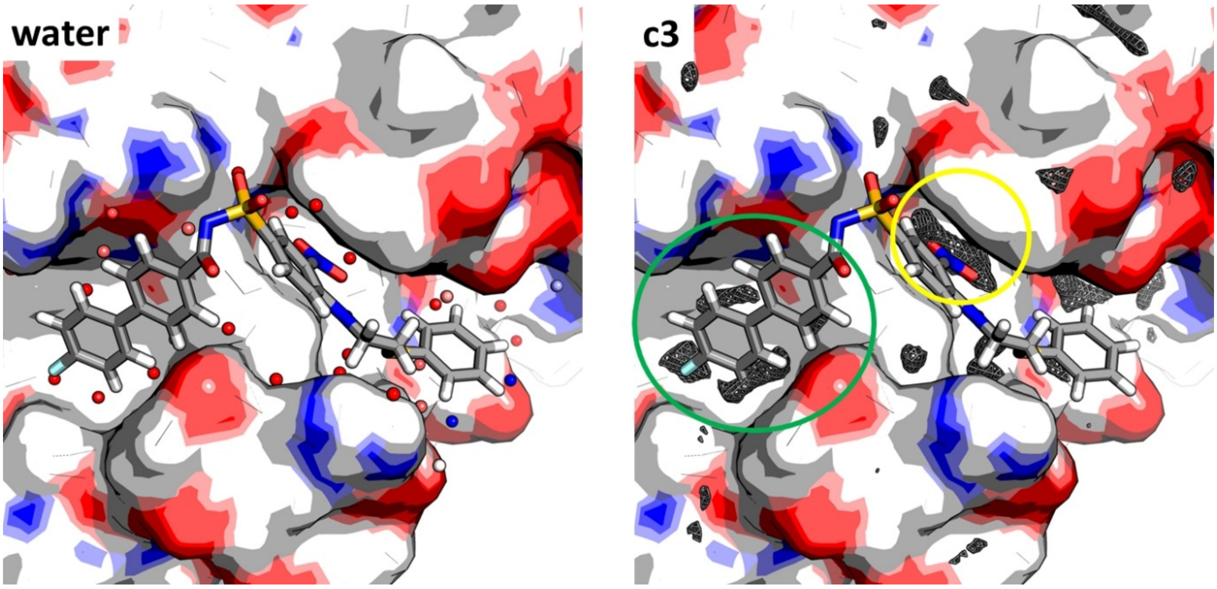

Figure 57: Close-up of the representative structure of the highest populated cluster for the MD of the Abbot inhibitor with predicted apo water molecules within 3.5 Å of the ligand coloured by their calculated $\Delta_{hyd}G_P$ contributions (left; from blue to red from -2.0 to +2.0 in units of kcal/mol) and the apo c3 probe density (right, threshold: 25). The biphenyl moiety corresponding to the carboxylic acid building block of the Ugi reaction is highlighted with a green circle in the right panel, the nitro group from the isonitrile building block in yellow. Surface colouring according to element identity. The respective raw data can be found in the electronic appendix (Electronic_Appendix/Bcl-xL/) in the respective structure folders (rep_1ysi).

According to the probe matching, this part of the ligand thus already perfectly matches the thermodynamic binding site profile, which might also be a reason for the high stability of this ligand part w.r.t. the simulations. The fluorine substituent, which is present in the Abbot inhibitor and Ugi model compound 1, coincides with a c3 peak maximum and an *apo* “unhappy” water molecule and thus can be considered to nicely match the properties of the surrounding pocket. Interestingly, regarding the results from 4.1.1, the respective part of the binding pocket represents another example where replacement of “unhappy” water molecules by an apolar ligand group is presumably highly favourable, as is also suggested by the matching of the water properties and the c3 probe density. Noteworthily, this correlation of “unhappy” waters and c3 probe density peaks in this part of the binding site can be observed for all investigated structures, both the crystal structure and docking poses and the representative MD structures. This again emphasises the large relevance that displacement of “unhappy” hydration sites by apolar ligand substituents can have for ligand binding.

W.r.t. library design, it is therefore advocated to use the biphenyl moiety with an apolar substituent in the *para*-position as a stability anchor with only minor modifications. Potentially, a polar group could be added in an *ortho*-position to the backbone amide: The probe analysis reveals high interpolated o and n4 *g*-function values in these *ortho*-positions for both structures of the Abbot inhibitor and for the docking poses of both Ugi model compounds. An Arg and Tyr sidechain in the proximity could potentially be targeted for ionic interactions or hydrogen bonds, for instance by adding a polar substituent or by introducing a heteroatom into the biphenyl system. In any case, the linker for use in DNA-encoded libraries should not be positioned in this part of the molecules since it would impede the stable binding in the hydrophobic groove.

4.3.2.2 Isonitrile part

The isonitrile building block introduces an aromatic ring system into the Ugi model compounds at which the phenylsulfanyl moiety is added via linker. In contrast to the Abbot inhibitors, this aromatic ring is not decorated with a nitro group in the Ugi model compounds.

In case of the Abbot inhibitor, the respective ring system adopts a similar orientation and location in the crystal structure and the representative MD structure (Figure 54). For the Ugi model compounds, the orientation and position of the phenyl ring show larger changes between the docking pose and the MD structure (Figure 53), which might be a hint that the respective part of the molecule is not as optimally bound as in the Abbot inhibitor.

Analysis of the water thermodynamics and probe densities in all used Bcl-xL structures shows that, similar to the biphenyl groove, the binding site region which accommodates the respective phenyl ring is rather hydrophobic and contains many high-energy water molecules. Especially the nitro group in the Abbot inhibitor shows high interpolated c3 probes *g*-function values and low values for the o and n4 probe. A close-up of the respective region in the representative cluster structure with the corresponding *apo* water molecules and the regions of maximum c3 probe density is shown in Figure 57. It reveals that the ring with the nitro group (yellow circle in Figure 57) is indeed located in a quite deep, hydrophobic pocket with two high energy *apo* water positions that coincide with the *holo* oxygen positions.

This is an interesting finding w.r.t. to further library design since the nitro group in the Abbot ligand was not introduced based on structural design but as a requirement for the synthesis route to modulate the reactivity of the ring system. The RISM-based analysis implies that this group is well suited for binding in this area in terms of steric properties but not in terms of chemical properties. Neither in the starting structure nor in the representative MD structure, suitable polar interaction partners are found in the proximity of the nitro group. Thus, substitution of the nitro group with a substituent which is similar in size but more apolar, for instance a methyl or halogen substituent, might still improve the ligand's affinity since it would better match the thermodynamic profile. As the nitro group is not necessary for the synthesis route of the Ugi compounds, this position offers a promising region for introducing modifications and improving the reduced affinity of the Ugi type compounds compared to the Abbot compounds. Due to the adaptivity of the binding site, it might even be possible to further exploit the respective pocket with larger groups. A search in the PDB revealed that there is already a published inhibitor which contains a SO_2CF_3 substituent in the respective position, with the near Tyr residue being flipped to accommodate the larger group (pdb 6qgj, Figure 58).

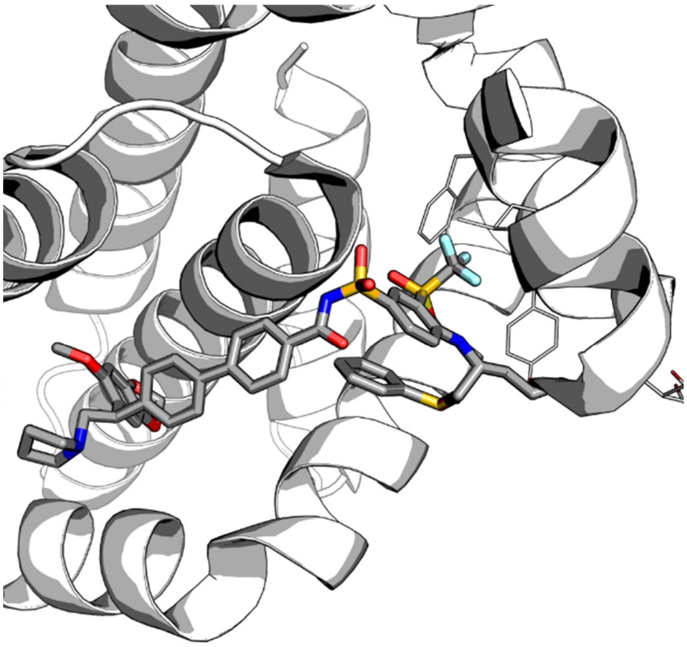

Figure 58: Bcl-xL crystal structure 6qgj with a modified substituent at the phenyl ring adjacent to the sulfone group.

W.r.t. linker placement, the respective ring system might be an option if sufficient fixation is achieved via a substituent. In the Abbot complex, the carbon atoms in *para* and *meta*-position to the nitro group point out of the binding site, thus offering possible positions for adding a linker chain.

4.3.2.3 Amine and aldehyde part

The R1 and R2 substituents in the Ugi backbone are introduced via the aldehyde and amine component of the Ugi reaction. In the two model compounds, only minimal substituents in form of H, Me, Et, or n-Pr were used in analogy to the Abbot inhibitor which does not contain corresponding substituents since the acylsulfonamide backbone is located in the respective region. However, both positions might thus offer possibilities to exploit areas of the binding site not accessible to acylsulfonamide-based compounds.

A close-up of the respective region the representative MD structures of both model compounds together with the c3, n4, and o probe density is shown in Figure 59. For both compounds, the respective R1 and R2 positions are oriented in a way that they point out of the binding site. This is promising w.r.t. linker placement since longer substituents could likely be added at both positions without hindering the overall binding of the molecule. In the representative structures, the substituents at the R2 position point towards a rather hydrophobic area on the surface of Bcl-xL which exhibits a lot of c3 probe density and minor o and n4 probe peaks at the rim of the area. Considering the adaptivity of the Bcl-xL binding site, this region could potentially be targeted by larger R2 substituents. This would allow to exploit a part of the protein surface which cannot be addressed by the Abbot-type inhibitors and thus is a promising strategy to improve binding affinity of the Ugi-based inhibitors. Therefore, the R1 position should be the preferred choice for addition of the linker, while a variety of larger substituents should be probed at the R2 position.

W.r.t. the backbone, it is noteworthy that, for model compound 2, the carbonyl oxygen atom nicely matches an o probe peak both in the docked and the representative structure. Although the predicted water molecules in this area exhibit rather high $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions (likely due to the overall rather apolar environment), the presence of an oxygen atom seems to be favourable in this area since it can undergo a hydrogen bond with a near Asn residue (Figure 54). Such a hydrogen bond between this Asn residue and a sulfone oxygen atom can also be observed in the representative MD structure of the Abbot compound. This indicates that the Ugi backbone is able to undergo stable, polar interactions with the Bcl-xL binding site similar to the Abbot compounds.

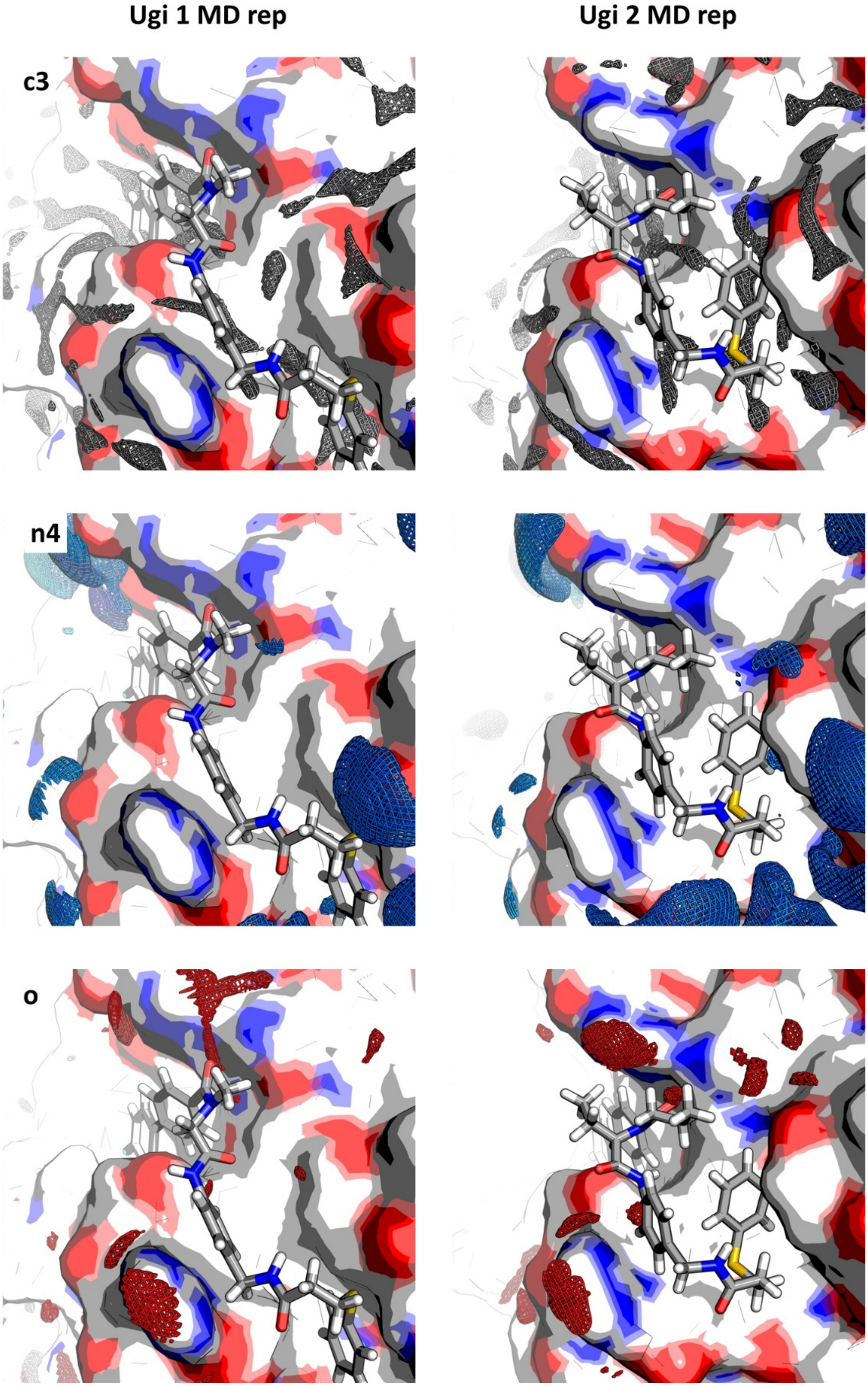

Figure 59: Close-up of the Ugi back bone part for the representative cluster structures of Ugi model compound 1 and 2 with apo c3 (threshold: 10), n4 (threshold: 200), and o probe (threshold: 10) densities of the binding site. The respective raw data can be found in the electronic appendix (Electronic_Appendix/Bcl-xL/) in the respective structure folders (rep_ugi_1, rep_ugi_2).

4.3.2.4 Amide coupling part

The part containing the phenylsulfanyl moiety is not introduced directly in the Ugi reaction but is added to the compound via a subsequent amide coupling. For both the Abbot and the Ugi-based compounds, this part exhibits by far the highest flexibility, implying that it does not undergo stable interactions. Together with the adaptive character of the Bcl-xL binding site, this makes rational design of this molecule part rather challenging. For the whole molecule part, there are no hydrogen bonds observed in any of the starting or representative MD structures. Introducing groups capable of forming stable polar interactions might therefore be a promising strategy for fixating this part of the molecule.

The interpolated probe densities (Figure 54 to Figure 56) in the respective molecule part show high n4 probe densities, suggesting that introduction of a polar or charged nitrogen group in this part of the molecule can serve as a stability anchor. Literature search revealed that there are indeed several published Bcl-xL inhibitors (2yxj, 3wiz, 4qvx, 6qgg, 6qgj) which contain a ternary amine in the respective region. Analysis on two representative examples (Figure 60, 2yxj, 4qvx) shows that the position of the amine nicely coincides with n4 probe peaks in the proximity of Asp residues in both cases. However, the two ligand parts have different shape and size, and the protein conformation and Asp side chain orientation also differ, again highlighting that the Bcl-xL binding site can accommodate groups of differing chain length, shape and size. It is therefore advocated to use this part of the molecule as a diversity region. Various substituents should be tested here, with polar or charged nitrogen groups in different positions to fully exploit the adaptivity of the binding site.

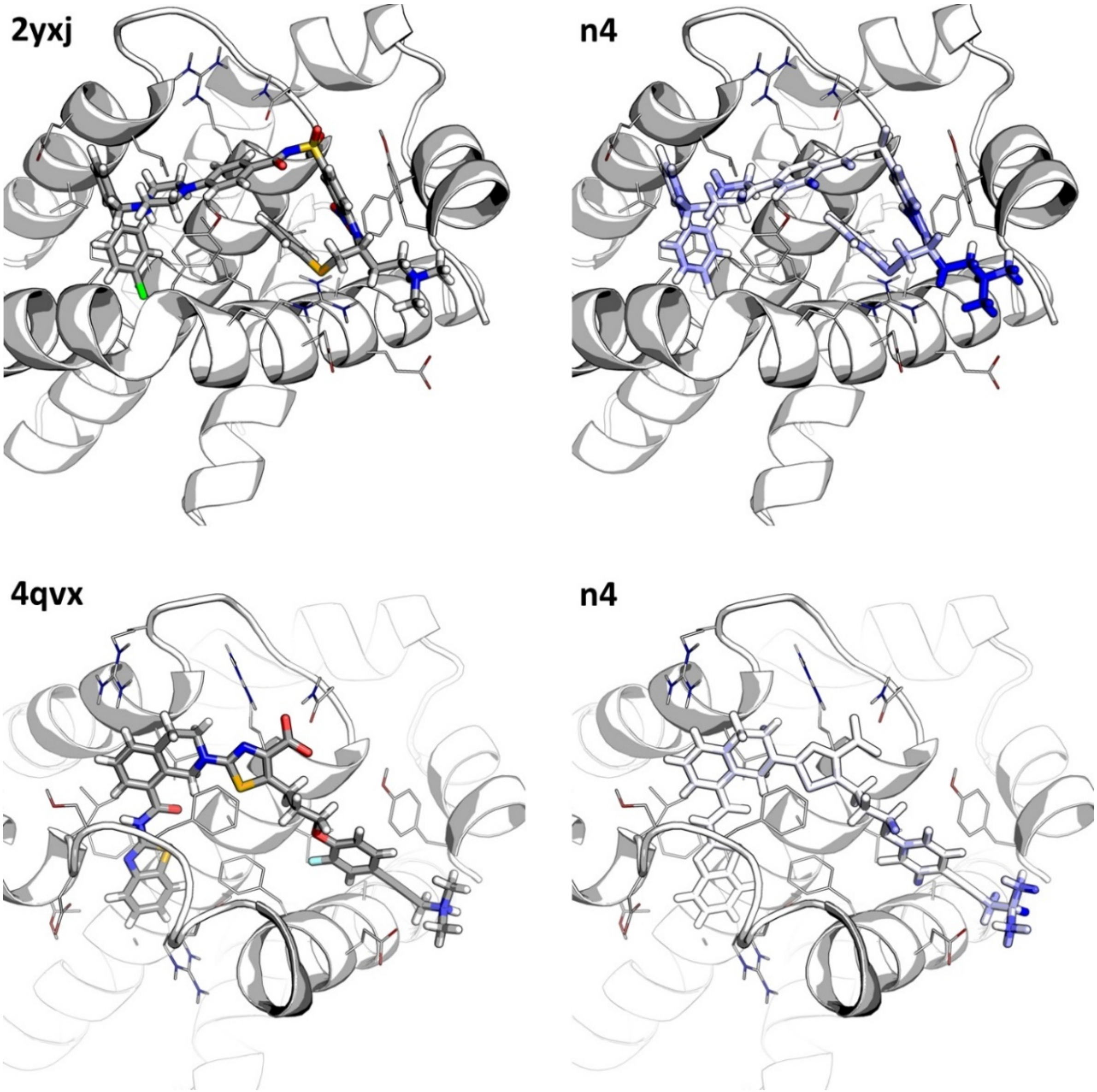

Figure 60: Structures and interpolated n4 probe density (colouring from white to blue from 0 to 50) for 2yxj and 4qvx. The respective raw data can be found in the Electronic Appendix (Electronic_Appendix/Bcl-xL/) in the respective structure folders (2yxj, 4qvx).

4.3.2.5 Summary of design proposals

The RISM-based binding site profiles for the Bcl-xL conformations tailored towards Ugi-based compounds yielded valuable information for the design of a respective combinatorial, DNA-encoded library. As summarised in Figure 61, it was found that the linker can be added at the R1 position via the aldehyde component without hampering the binding of the molecule. The biphenyl moiety, as introduced by the carboxylic acid component, should be used as a stability anchor, potentially with the introduction of polar groups in the ring system adjacent to the backbone. In the isonitrile moiety, an apolar substituent should be introduced to exploit the apolar pocket which, in the complex with the

Abbot inhibitor, accommodates the nitro group. Diversity can be introduced at the R2 position via the amine component of the Ugi-reaction and via the eastern part of the molecule which is added to the isonitrile part via amide coupling. The group at the R2 position could target so far unexploited areas on the surface of Bcl-xL by for instance addressing Tyr159. In the amide coupling part, a large variety of substituents can be introduced. They should however contain polar or charged nitrogen groups to target the region containing two Asp residues that could be exploited for forming hydrogen bonds or a salt bridge to fixate the molecule in the binding site.

While some of these results could also have been derived from the analysis of the original crystal structure 1ysi alone (good matching of the biphenyl moiety, substitution proposal for the isonitrile part), the inclusion of representative structures from respective MD simulations added valuable insights w.r.t conformational stability. For instance, comparison between the starting structures and the representative structures confirmed the high stability of the biphenyl moiety and the flexibility of the amide coupling part. In the future, however, the proposals made here have to be validated by experimental data.

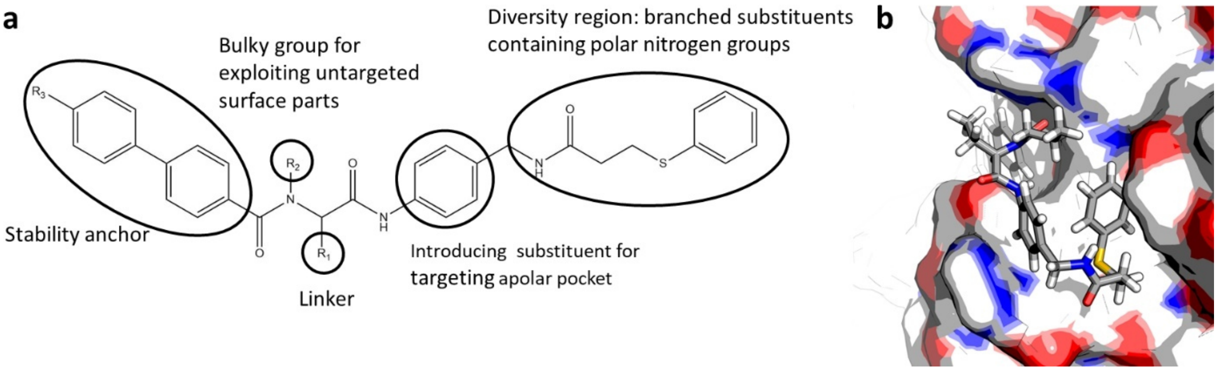
a
Diversity region: branched substituents containing polar nitrogen groups
Bulky group for exploiting untargeted surface parts
R_3
R_2
O
N
H
S
Stability anchor
R_1
Introducing substituent for targeting apolar pocket
Linker
b

Figure 61: a) Design strategy for the Ugi-reaction based combinatorial DNA-encoded library tailored towards Bcl-xL. b) Representative structure of the highest populated cluster for the MD of Ugi model compound 2.

4.3.3 TEAD – Investigating large protein-protein-interfaces

The interaction between the transcriptional enhancer factor domain (TEAD) and the co-transcription factor Yes-associated protein (YAP) plays an important role in the Hippo pathway which was found to be dysregulated in many types of cancer.[314] Disrupting this interaction is therefore an attractive potential therapeutic strategy. However, the PPI is a challenging target since the interface is large and spans over different regions, involving dozens of amino acids (Figure 62), so that so far only few small molecule inhibitors are known.

Via screening of DNA-encoded indole-focused Ugi peptidomimetics, the Brunschweiger group could identify compounds containing a Cl-indole moiety that inhibit the interaction of human TEAD4 with YAP with IC_{50} values in the low micromolar range.[314] The aim of the presented study therefore was to derive reasonable predictions where and how the respective compounds might bind, and to derive a design strategy for another indole-focused DNA-encoded library that allows for synthesis via click reaction. Therefore, at first, the *apo* binding site of TEAD as well as complexes with known peptide binding partners were characterised to elucidate respective SAR trends. Based on the gained insights, constrained docking studies of model compounds for a DNA-encoded library using click reaction provided by the Brunschweiger group were carried out to predict the potential binding modes and to make suggestions for building block selection.

4.3.3.1 Characterizing the TEAD binding site and interactions with peptides

TEAD interacts with YAP at two distinct sites on its surface, one binding amino acids 61-73, which exhibit a helical conformation, and one binding the amino acids 85-99, which exhibit a Ω-loop conformation.[315] An overview of the respective interaction and a close-up of the Ω-loop region are shown in Figure 62.

A hit fragment identified from experimental screening was shown to bind in a pocket where Phe69 of YAP is accommodated in complex with hTEAD4.[316] In addition, TEAD also contains a central pocket where it is palmitoylated and where several small molecules are known to bind.[314] This large number of potential interaction sites makes it hard to make a reasonable prediction where and how the indole-containing hit compounds by Kunig *et al*.[314] might bind. The natural YAP peptide does not contain any Trp residues but several other apolar residues like Phe, Met and Ile.

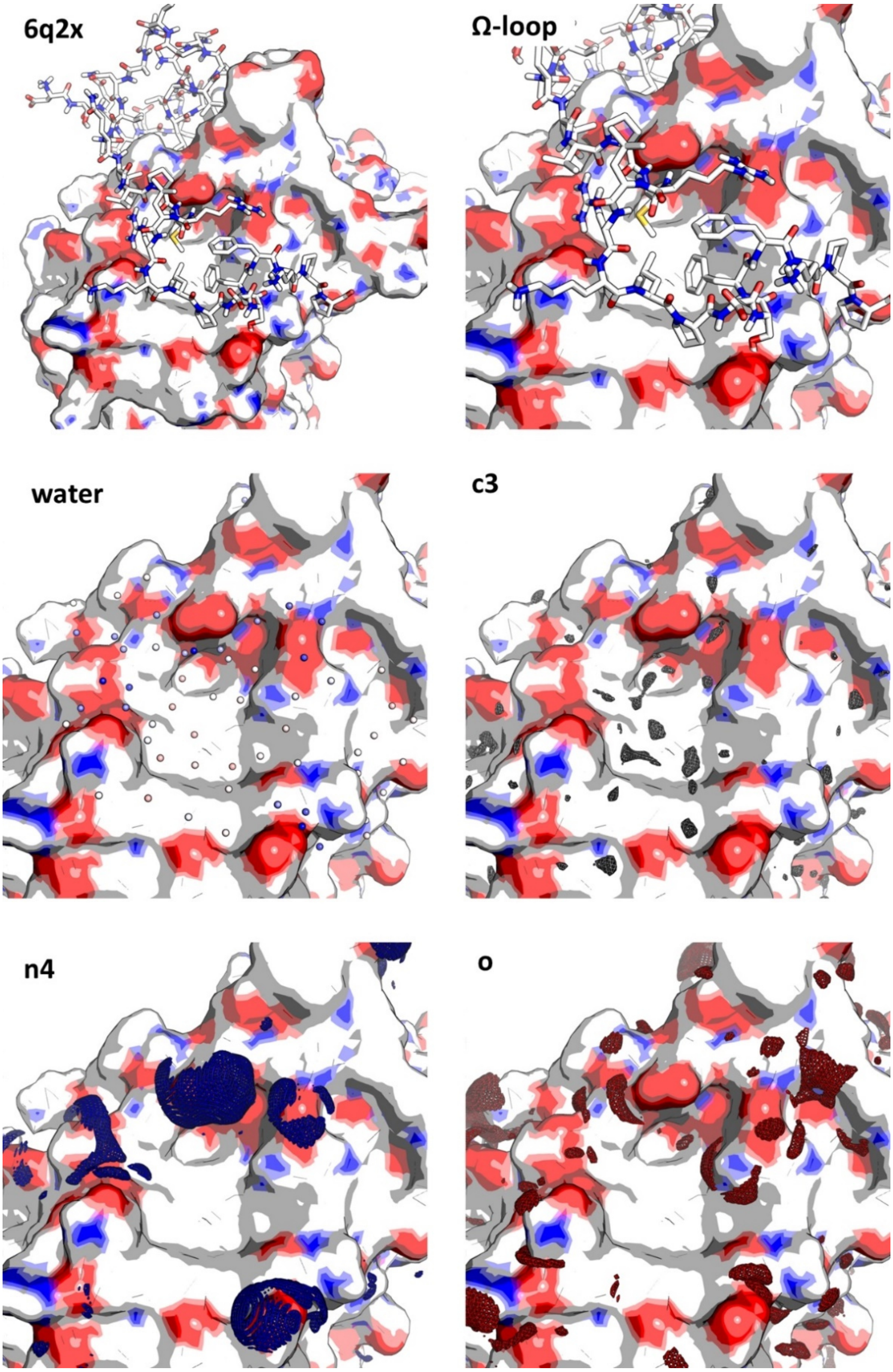

Figure 62: Binding site characterisation of hTEAD4 based on complex structure 6q2x (TEAD4-YAP interaction) with close-up of the Ω-loop region. Predicted apo water sites (colouring according to $\Delta_{hyd}G_P$ contributions from blue to red from -2.0 to +2.0 in units of kcal/mol) as well as c3, n4, and o probe densities are shown (thresholds: c3 - 25; n4 - 100; o - 40). The respective raw data can be found in the electronic appendix (Electronic Appendix/ TEAD/ 6q2x).

However, fortunately, Furet *et al*. performed structure-based design of several peptides that bind at the Ω-loop region and inhibit the YAP-TEAD interaction in the nanomolar range.[315] They achieved this by introducing unnatural amino acids, including the mutation of Met86 to Trp and 6-Cl-Trp, which led to an approx. 3- and 12-fold increase in potency compared to the original peptide. Due to the massive improvement in potency upon introduction of 6-Cl-Trp, it is reasonable to assume that the screening hits by Kunig *et al*. are likely to bind in the same region via the Cl-indole moiety.

To further investigate the respective binding site and draw conclusions about binding mode and SAR, RISM-based calculations of the binding site water molecules and pharmacophoric probe densities were performed on both the *apo* TEAD4 protein and the YAP-TEAD4 complex (Figure 62).

As shown in Figure 62, the respective Ω-loop region is located in a hydrophobic groove surrounded by polar residues. As a consequence, the *apo* site contains several high energy water sites and c3 probe density peaks, whereas o and n4 probe density is found at the rim of the region. In the natural YAP-TEAD4 complex, Met86, Leu91, and Phe95 are accommodated in the hydrophobic groove. However, when studying this cavity in detail (Figure 63), as Furet *et al.* did for structure-based design of the peptides,[315] it is apparent that especially the Met residue does not fully exploit the cavity. Results of 3D RISM-based calculations on the YAP-TEAD4 complex show that, in the presence of YAP, there is still considerable c3 probe density in the deeper regions of the groove as well as multiple high energy water sites which are trapped in the remaining space (Figure 63).

Overlay with the structure including 6-Cl-Trp shows that especially the chlorine overlaps with the biggest c3 probe peak in the region and with a high energy water site which thus is replaced in the complex with the modified peptide (Figure 63). This release of a high energy water molecule might also explain the large gain in affinity when switching from Trp to Cl-Trp. This, again, is a strong indication that the Met86 cavity is indeed the anchor region for the respective peptidomimetics.

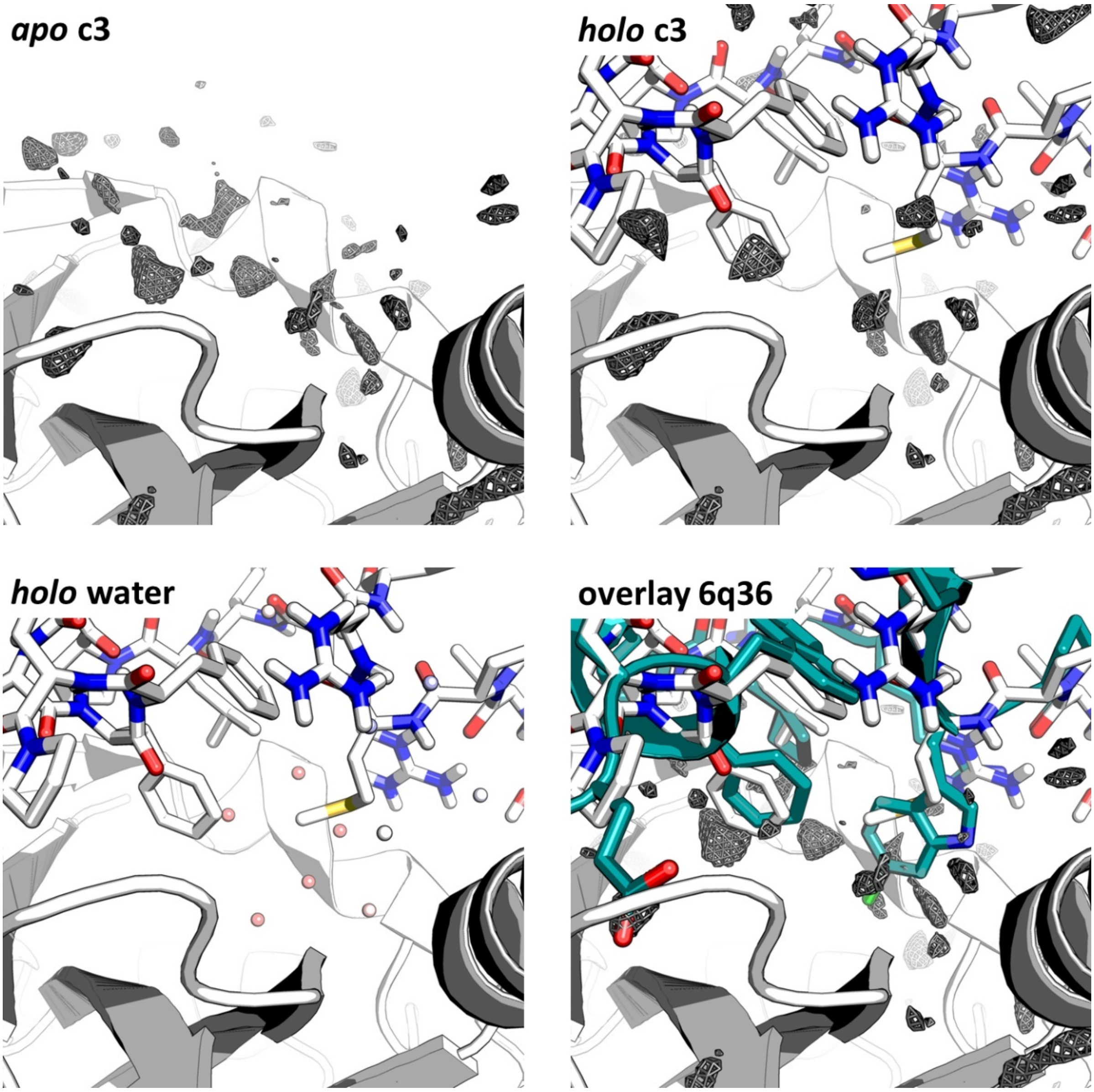

Figure 63: Apolar cavity in the Ω-loop region of the YAP-TEAD4 interaction in structure 6q2x. Upper left: apo c3 probe density (threshold: 25); upper right: holo c3 probe density (threshold: 25); bottom left: holo water thermodynamics (colouring according to $\Delta_{hyd}G_P$ contributions from blue to red from -2.0 to +2.0 in units of kcal/mol); bottom right: overlay with the modified peptide in 6q36 (green) and apo c3 probe density (threshold: 25). The respective raw data can be found in the electronic appendix (Electronic Appendix/ TEAD/ 6q2x).

Encouraged by the agreement of this thermodynamic analysis and the SAR trends observed by Furet *et al*., two other sites were analysed where modifications were introduced. Pro92 was modified to Hyp to form a hydrogen bond with a backbone carbonyl (Figure 65). The RISM-based analysis shows that the hydroxyl group replaces a water molecule with average thermodynamic properties in a region where a small o probe density peak can be observed. Here, the affinity gain thus likely originates directly from the hydrogen bond and not from the release of a water molecule.

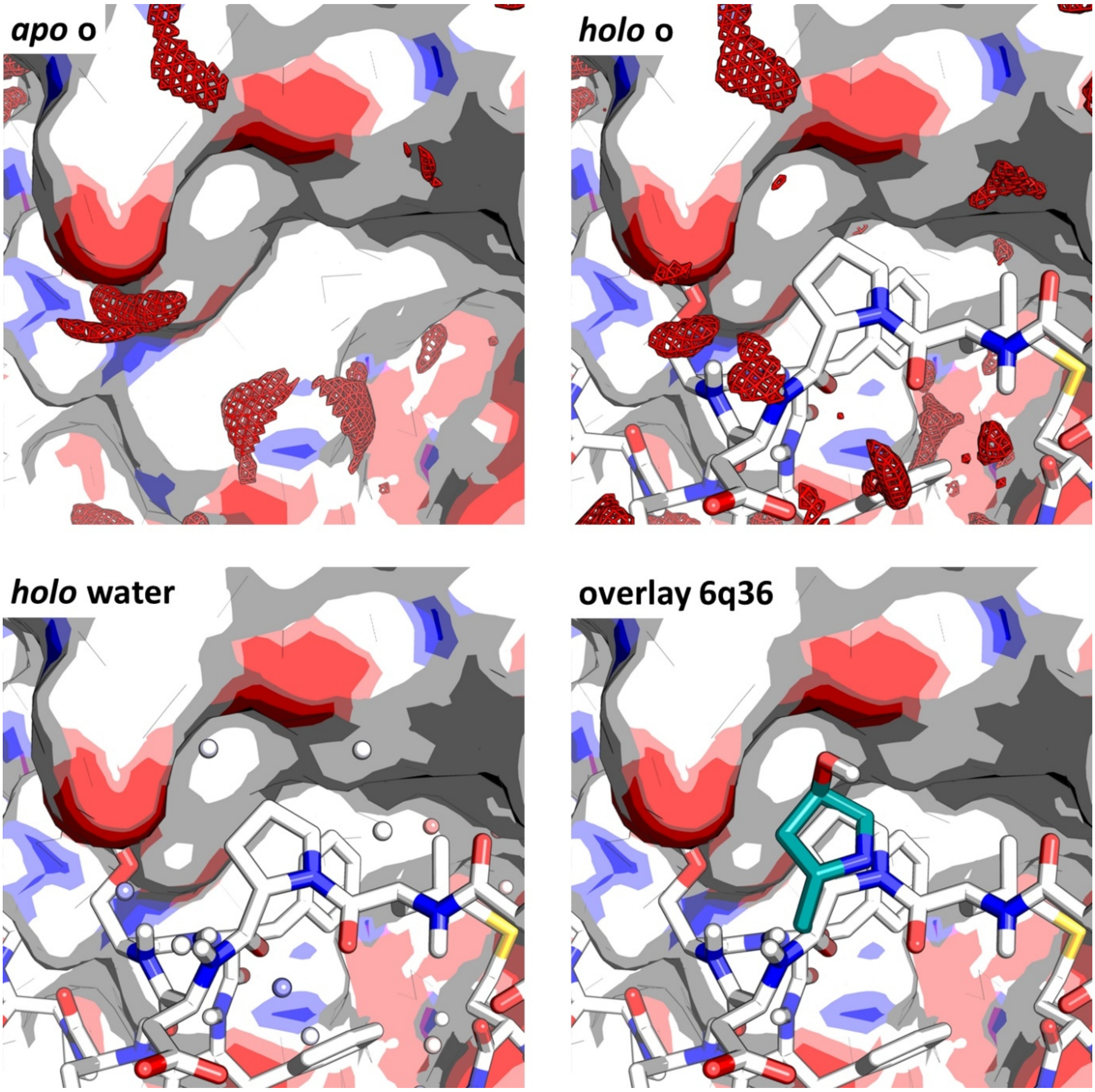

Figure 64: Pro to Hyp modification from 6q2x to 6q36. Upper left: apo o probe density (threshold: 40); upper right: holo o probe density (threshold: 40); bottom left: holo water thermodynamics (colouring according to $\Delta_{hyd}G_P$ contributions from blue to red from -2.0 to +2.0 in units of kcal/mol); bottom right: overlay with the modified Hyp residue in 6q36 (green). The respective raw data can be found in the electronic appendix (Electronic Appendix/ TEAD/) in the respective pdb folders.

Another modification is the substitution of Pro98 to a Glu residue (Figure 65). Here, the newly introduced carboxyl group occupies an area with high o probe density in both the *apo* TEAD structure and the *holo* TEAD-YAP complex. In addition, the oxygen atoms coincide with the positions of two "unhappy" water molecules in the TEAD-YAP complex. This implies that addition of the carboxylic group in this area is favourable both because of the improved interaction with the TEAD binding site residues and because of the release of the respective high energy water molecules.

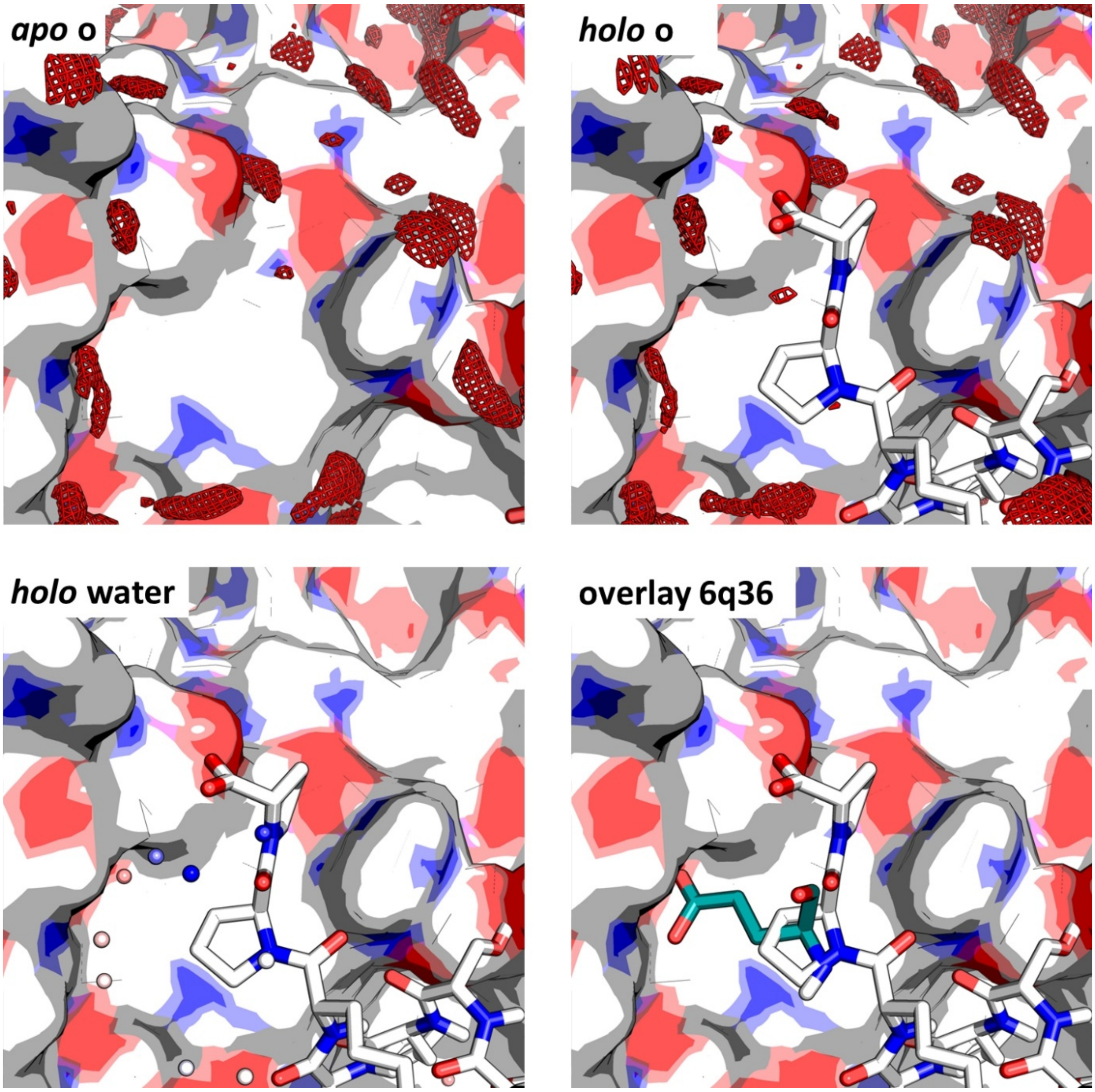

Figure 65: Pro to Glu modification from 6q2x to 6q36. Upper left: apo o probe density (threshold: 40); upper right: holo o probe density (threshold: 40); bottom left: holo water thermodynamics (colouring according to $\Delta_{hyd}G_P$ contributions from blue to red from -2.0 to +2.0 in units of kcal/mol); bottom right: overlay with the Glu residue in 6q36 (green). The respective raw data can be found in the electronic appendix (Electronic Appendix/ TEAD/) in the respective pdb folders.

All in all, the retrospective study on the modified peptide thus show that the water thermodynamics and probe densities are well-suited to explain the respective SAR trends, suggesting that their combination with docking studies should be a promising approach to elucidate potential binding modes of the indole-containing model compounds and to propose further strategies for a screening library design.

4.3.3.2 Docking of peptidomimetics

To guide the selection of Cl-indole containing compounds for a DNA-encoded library, a set of 48 compounds as proposed by the Brunschweiger group (list in Appendix, 7.4) was docked into the TEAD4 structure 6q36 (poses can be found in the Electronic Appendix (Electronic Appendix/TEAD/Docking/)). The core structure of the compounds in the data set and the structures of used substituents are shown in Figure 66. All molecules contain the 6-Cl-indole moiety which is also present in the modified peptide of Furet *et al.*,[315] with an exit vector at the 2-position containing a triazole moiety, thus allowing for synthesis via click reaction.

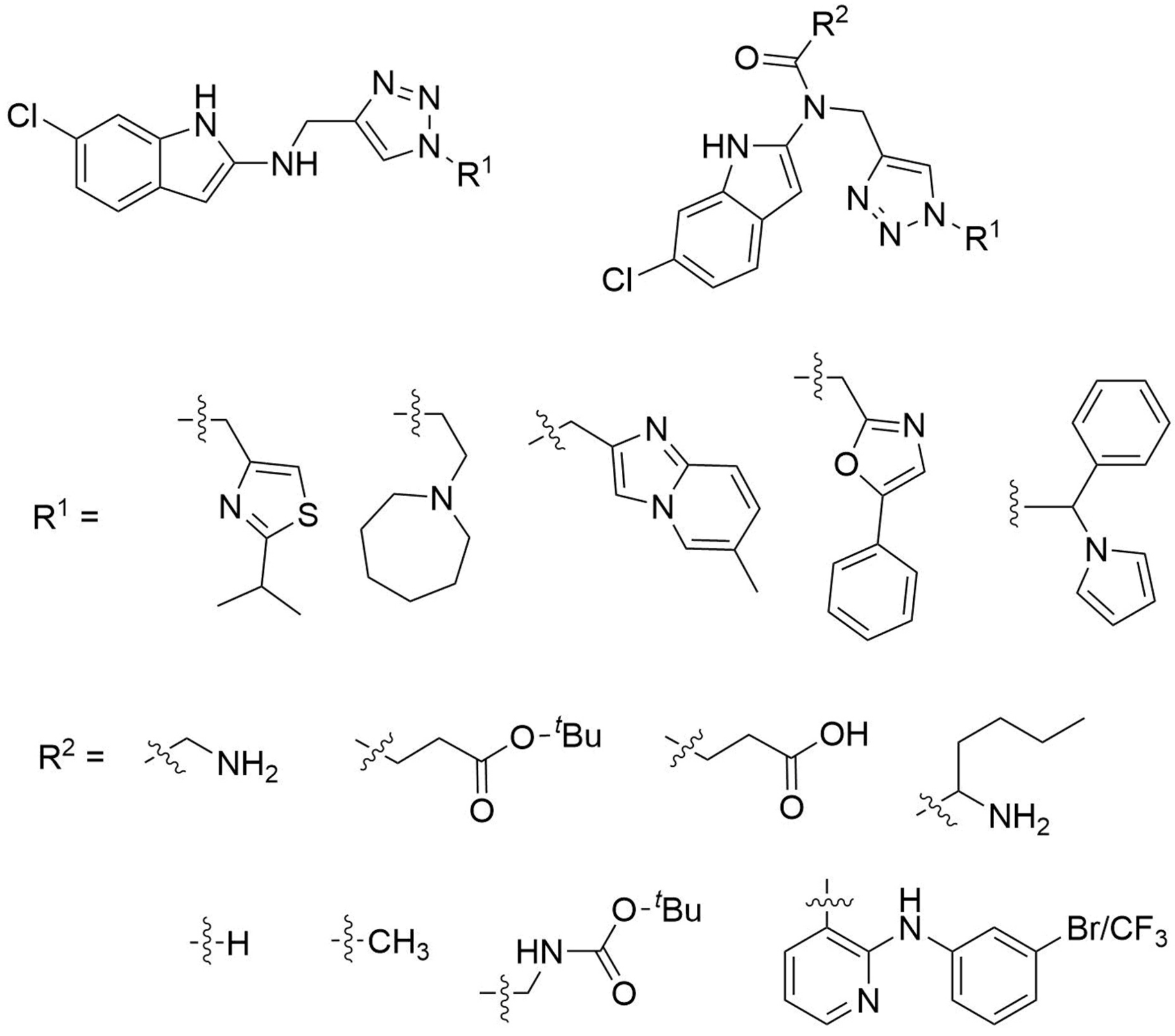

Figure 66: Core Structures and used substituents of the compounds in the TEAD4 data set provided by the Brunschweiger group.

Since the 6-Cl-indole moiety was shown to be a key factor for the enhanced affinity of the modified peptide, it was used as a constrained anchor position for the docking experiments. The resulting poses were then then evaluated w.r.t. a consensus ranking based of the summed ranks according to conventional ChemPLP score, a normed ChemPLP score per heavy atom, and the ligand-probe matching score as presented in 4.2. This way of ranking was chosen to achieve an optimal balance between the scoring metrics and to prevent the favouritism of especially large or small compounds in the data set.

Figure 67 shows an overview of the binding site in the respective structure 6q36 and an overlay of the bound peptide with the docking pose of the top ranked molecule for clarity. The docking poses of the six top ranked molecules are shown in Figure 68 in the same orientation as in Figure 67. The corresponding scores and ranks of all molecules are listed in Table 26.

It is noteworthy that, with exception of rank 2, the predicted binding modes are in good agreement with each other, with the triazole-linked branch of the molecule pointing out of the hydrophobic groove flanked by two Glu residues, while the second branch is located near Lys273. This was also observed for most of the other molecules (/Electronic_Appendix/ TEAD/ Docking/). Considering the rather large number of rotatable bonds in the molecules and the large binding site, this is a promising finding w.r.t. the reliability of the docking experiment.

Interestingly, as can be seen in the overlay of the top ranked pose with the peptide from 6q36 (Figure 67), the docking predicts the molecules' triazole-linked branch to occupy a region of the binding site that is not exploited by the peptide, while the branch containing the R2 substituent is predicted to lie in the same area as a Pro residue in YAP. The large hydrophobic groove that accommodates the Phe and Cba residues of the peptide in 6q36 is not occupied in the respective docking poses, which however seems reasonable due to the rather polar character of the compounds.

W.r.t. the R2 substituent, the consensus scoring shows a strong preference for an amino group: Four of the six top ranked compounds, rank 1, 3, 4, and 5, contain the amino substituent, and in all cases it is predicted to lie at the polar rim of the hydrophobic pocket near Glu391 so that a salt bridge can be formed. In the compound on rank 6, the R2 substituent is a succinate-tBu-ester which is predicted to lie in the area where the terminal Pro residue of the modified YAP-peptide is located in 6q36. Due to the length of the succinate chain, the ester group extends towards Lys273 and can thus form a hydrogen bond via the carbonyl oxygen atom. The respective ester was also present in the hit compounds by Kunig *et al*. which however contained a different core structure.[314]

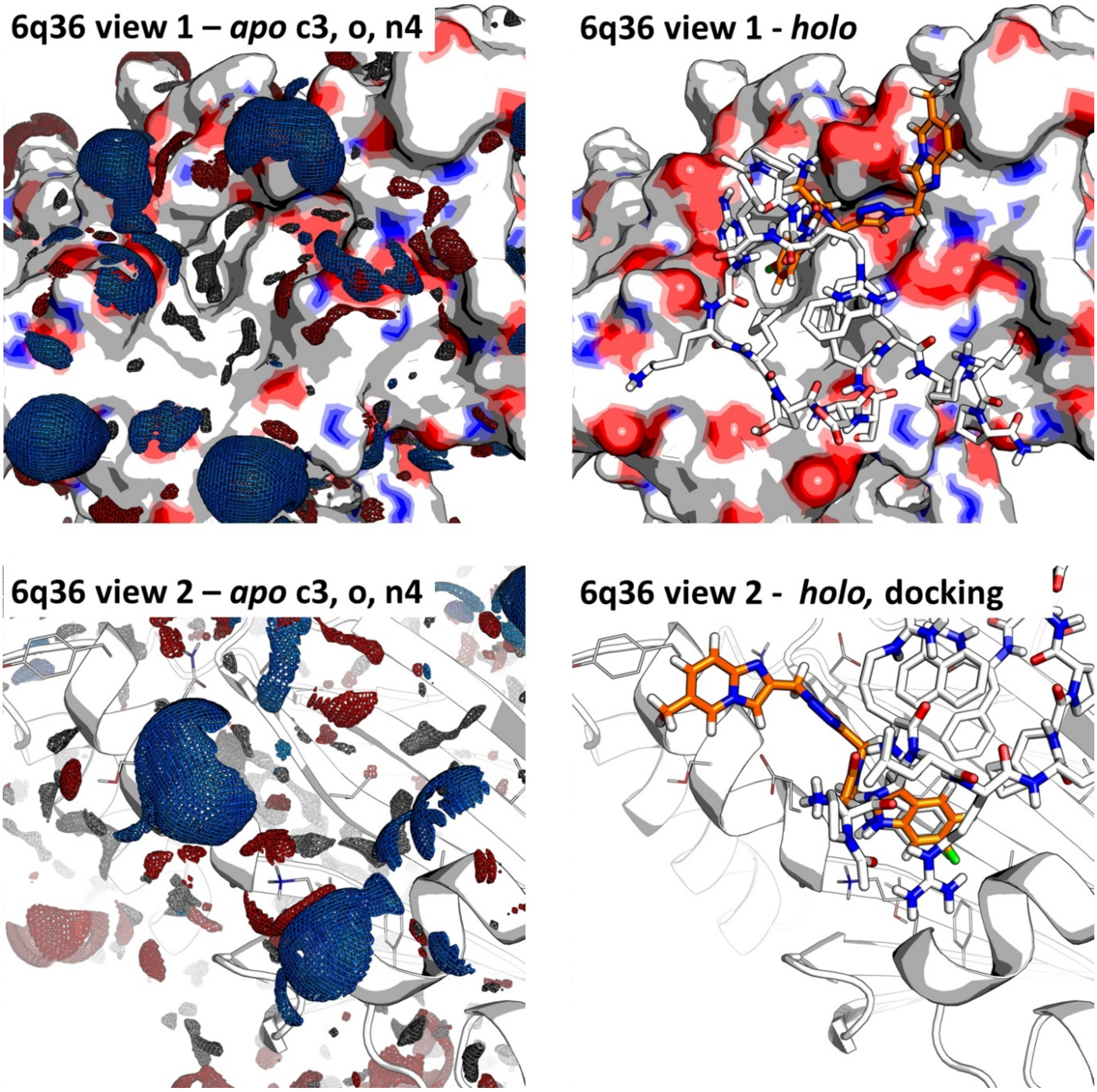
6q36 view 1 – *apo* c3, o, n4
6q36 view 1 - *holo*
6q36 view 2 – *apo* c3, o, n4
6q36 view 2 - *holo,* docking

Figure 67: Apo probe densities in 6q36 (thresholds: c3 - 25; n4 - 100; o - 40) (upper left and bottom left) and overlay of the respective bound peptide with the top ranked docking pose for the Brunschweiger compounds (upper right and bottom right). The respective raw data can be found in the electronic appendix (Electronic_Appendix/ TEAD/ 6q36/ and /Electronic_Appendix/ TEAD/ Docking/).

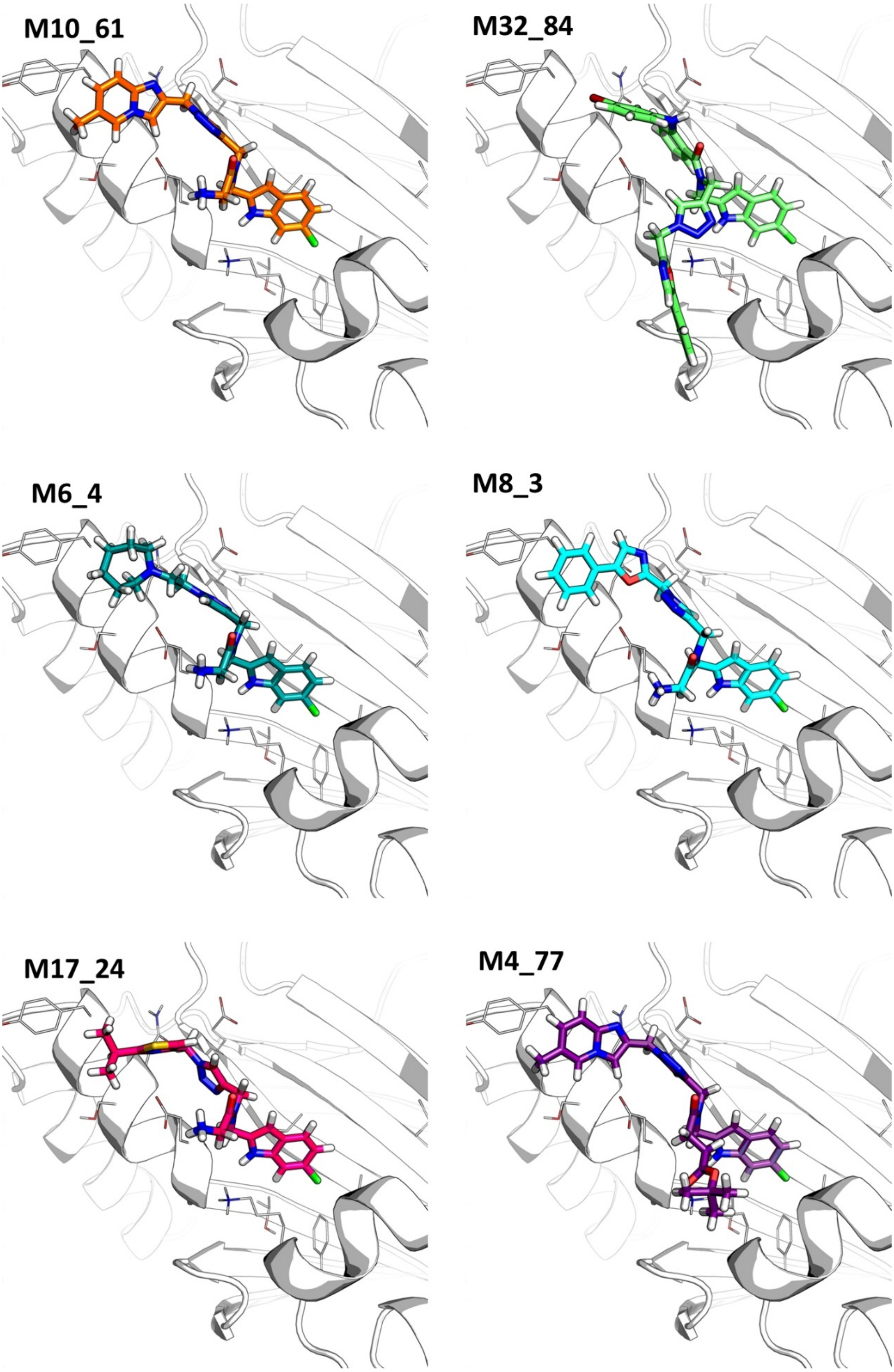

Figure 68: Docking poses of the six top ranked molecules of the TEAD4 data set provided by the Brunschweiger group in the binding site of 6q36. The respective raw data can be found in the Electronic Appendix (/Electronic_Appendix/ TEAD/ Docking/).

Table 26: Scores and corresponding ranks for the docked compounds provided by the Brunschweiger group according to conventional ChemPLP score, normalized ChemPLP score, the ligand-probe matching score s_{pose}, and a resulting consensus rank (generated by summing up the ranks of all three individual scores and sorting in ascending order). The respective raw data (poses and scores) can be found in the electronic appendix (/Electronic_Appendix/ TEAD/ Docking/). A plot of the conventional ChemPLP scores and normalized ChemPLP scores against the ligand-probe matching score s_{pose} is provided in the Appendix (7.8).

Cmpd	ChemPLP		ChemPLP/*N*(HA)		s_{pose}		consensus
	score	rank	score	rank	score	rank	rank
m1_71	39.71	48	2.65	1	0.231	9	13
m10_61	70.14	18	2.13	4	0.319	4	1
m11_14	70.00	19	1.94	13	0.192	21	9
m12_91	56.33	46	2.09	8	0.197	18	24
m13_29	63.18	37	2.18	2	0.227	10	7
m14_94	56.33	47	1.48	43	0.135	41	48
m15_76	58.82	42	1.96	12	0.586	1	10
m16_90	64.68	33	1.70	29	0.155	34	42
m17_24	64.80	31	2.09	6	0.245	7	5
m18_83	65.21	28	1.86	15	0.115	45	35
m19_11	62.31	38	1.64	33	0.200	16	34
m2_20	58.39	43	2.16	3	0.183	25	23
m20_59	67.34	25	1.37	47	0.174	28	43
m21_95	65.10	29	1.59	37	0.141	39	44
m22_54	74.30	9	1.52	41	0.162	29	30
m23_2	63.19	36	1.50	42	0.091	48	47
m24_35	60.70	41	2.02	9	0.217	12	16
m25_93	74.10	10	1.61	36	0.108	46	39
m26_31	63.81	34	1.39	46	0.213	13	40
m27_43	73.13	12	1.56	39	0.138	40	38
m28_82	84.48	1	1.54	40	0.200	17	14
m29_32	64.85	30	1.47	44	0.226	11	33
m3_38	65.39	27	1.72	27	0.145	35	36
m30_43	69.33	21	1.58	38	0.202	15	25

Cmpd	ChemPLP		ChemPLP/*N*(HA)		s_{pose}		consensus
	score	rank	score	rank	score	rank	rank
m31_6	76.88	5	1.79	18	0.105	47	21
m32_84	82.70	2	1.80	17	0.243	8	2
m33_9	75.47	8	1.68	32	0.157	30	22
m34_78	72.49	13	1.69	30	0.156	33	27
m35_3	58.29	44	1.94	14	0.182	26	32
m36_29	73.22	11	1.79	19	0.191	22	8
m37_12	76.37	6	1.74	25	0.143	37	19
m38_93	72.15	14	1.68	31	0.157	31	28
m39_15	79.60	4	1.73	26	0.141	38	20
m4_77	69.60	20	1.74	23	0.322	3	6
m40_78	79.65	3	1.63	35	0.120	44	31
m41_80	75.61	7	1.84	16	0.156	32	11
m42_87	67.78	24	1.41	45	0.249	5	26
m43_52	61.00	40	1.33	48	0.486	2	37
m44_7	63.46	35	1.98	11	0.194	19	18
m45_57	57.40	45	1.74	24	0.133	42	46
m46_58	61.83	39	1.63	34	0.143	36	45
m47_22	68.62	22	1.72	28	0.126	43	41
m48_86	72.12	15	1.76	20	0.178	27	17
m5_42	71.46	16	1.74	22	0.194	20	15
m6_4	65.62	26	2.12	5	0.245	6	3
m7_70	68.34	23	2.01	10	0.188	23	12
m8_3	71.07	17	2.09	7	0.208	14	4
m9_90	64.73	32	1.75	21	0.185	24	29

Thus, the consensus scoring clearly advocates the use of polar substituents for R2 to target the charged residues Lys273 and Glu391 at the rim of the binding site. This might explain the flipped binding mode which is predicted for the compound on rank 2: Here, the R2 substituent is a rather hydrophobic niflumic acid moiety which could not undergo respective polar interactions. Rather, it is predicted to extend over the rim towards another hydrophobic area, so that instead the R1 branch is positioned near Glu391 and Lys273.

W.r.t. to the R1 substituents, the consensus scoring shows a slight preference for the imidazopyridine, which was also present in the initial hits by Kunig *et al*.[314] Like the niflumic acid moiety in the rank 2 molecule, it extends over the rim to a hydrophobic area where the methyl substituent nicely coincides with a c3 probe density peak (Figure 67). A similar binding mode is predicted for the thiazole substituent bearing an isopropyl group, for the oxazole substituent with a phenyl group, and for the azepane substituent. In case of the latter, the positively charged amine lies in an area with high n4 probe density between Glu416 and Glu391, thus potentially allowing the formation of salt bridges.

Thus, the presented study was able to propose a reasonable, potential binding mode for the molecules in the respective library, using the assumption that the Cl-indole moiety of the respective compounds binds in the same region as the Cl-indole of the modified YAP peptide. Given the predicted poses, the consensus scoring suggests to combine polar R2 substituents with rather hydrophobic R1 substituents to build compounds which can optimally target the area around the polar rim of the YAP binding site. For targeting the hydrophobic groove of the binding site, which is occupied by apolar residues of the YAP peptide, additional exit vectors or substituents at the indole moiety could be added.

Although detailed experimental affinity data and crystal structures will be needed to validate these predictions, the presented case study showed how the RISM-based water analysis and probe densities can be used to explain SAR trends for challenging PPI targets and to guide the design and synthesis of suitable screening candidates.

5. Summary and outlook

In this work, novel algorithms and workflows were developed that now allow to directly convert primary results of 3D RISM calculations into interpretable parameters – and ultimately a score – to guide the structure-based design of a ligand for a given protein target.

The first part of this work focused on the thermodynamic properties of binding site water molecules and their influence on ligand binding and ligand affinity. Building on previous work within the Kast group, a framework was developed that does not only permit to determine the $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions of specific, predicted hydration sites but to directly map *apo* water thermodynamics onto atoms of bound ligands. This is a very powerful tool for SBDD since it opens up the possibility to directly correlate binding site water thermodynamics with structural ligand features and even with ligand affinity, which was not possible in such a direct and convenient way before.

To ensure a meaningful and statistically significant analysis, the developed approaches were applied onto a large data set comprising several thousand protein ligand complex structures – an order of magnitude that has, to the best knowledge of the author, not yet been covered in published studies using any other approaches related to water thermodynamics in binding sites. A key finding from the analysis in this work was that the thermodynamic properties of a given hydration site are coined by the precise microenvironment and cannot be directly connected to experimental parameters like B-factors. Hence, approaches as presented in this work are needed to gain such information.

A key part of this work then was to correlate *apo* water thermodynamics with ligand replacement. Water molecules were characterised as "happy" or "unhappy" depending on whether they exhibit a favourable or unfavourable contribution to the total free energy of hydration of the protein, $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$, as calculated by the 3D-RISM based algorithms. The analysis revealed that, although "happy" and "unhappy" water molecules are replaced alike within the whole data set, replacement of more "unhappy" hydration sites favourably correlates with ligand binding affinity and druggability. Following this result, replacement preferences for different elements and atom types were investigated. By then correlating these found trends with provided binding affinity data, a hint towards a hidden bias in the data set was revealed: The correlation with ligand affinity showed that replacement of a water molecule with an oxygen atom is not always optimal but rather depends on the specific thermodynamic properties of the targeted hydration site. For instance, hydroxyl groups should be used to target "happy" water molecules while apolar substituents should be used to replace "unhappy" water molecules. To illustrate the findings from the large-scale analysis, several case studies of MMPs were presented for which both the *apo* and *holo* water thermodynamics can be used to explain – at least to a certain extent – the affinity differences between the closely related molecules. In the future, the "replacement rules" revealed in this work can be employed to lead the rational design and optimisation of ligands for any given protein target.

Apart from the direct impact on rational drug design, the presented study also impressively demonstrated that some trends derived from large-scale data might actually not be attributed to physical principles but to bias introduced by common design principles. Thus, it also highlights the need for the critical assessment of available data to gain an optimal benefit for future research.

In the second part of the work, the concept of local thermodynamic binding site characterisation was extended towards pharmacophoric probes mimicking ligand functional groups. With the implementation of the 3D RISM *uu* formalism by F. Mrugalla as described in 2.3.3.2, the distribution of those probes in the binding site can be determined. As for the water thermodynamics, an approach was presented to directly map the *apo* *g*-function values of these probe onto atoms of the bound ligand, thus allowing to analyse if areas with high *g*-function values of a given probe coincide with the presence of corresponding ligand atoms. An analysis on the protein-ligand complex structures in the PDBbind core set indeed revealed a good matching between probe density peaks in the binding site and the presence of respective ligand groups. Building on this, a quantitative ligand-probe matching score was developed to capture how well a bound ligand in a given complex (or docking pose) matches the pharmacophoric probe distribution. In a proof-of-concept study w.r.t. pose recovery docking using the PDBbind core set it was shown that the score indeed correlated with the RMSD of the poses, i.e. that poses with higher ligand-probe matching scores were closer to the native binding mode. After this encouraging analysis, a scoring scheme using a combination of conventional ChemPLP score and the ligand-probe matching score was applied in a proof-of-concept study of virtual screening on a respective data set for XIAP. Here, the ROC AUC, i.e. the retrieval of actives before decoys, could be strongly improved by applying the ligand-probe match score as a filtering criterion before ranking by conventional ChemPLP score. This strategy is especially promising for virtual screening of large libraries since inactive molecules can be filtered out effectively.

Another highly promising application of the ligand-probe matching score is the identification of suitable molecule fragments that can be used as a starting point for ligand design or could serve as building blocks for the synthesis of combinatorial libraries. W.r.t. to the collaboration with the Brunschweiger group, this application was of special interest for this work. Therefore, a semi-automated workflow was developed comprising automated library preparation in KNIME and docking in GOLD, followed by scoring according to the ligand-probe matching score. A proof-of-concept study on XIAP showed that a combination of normalised ChemPLP score and the ligand-probe matching score is well suited to retrieve known binders from the rest of the fragments. Hence, the developed workflow can now be employed to identify suitable molecule fragments that ideally match the respective binding site profile of a given target - thus making one more step towards the automated *de novo* ligand design.

In the third part, the water analysis and the pharmacophoric probes combined were used to tackle challenging PPI targets in medicinal chemistry that are under investigation in the Brunschweiger group: XIAP, Bcl-xL and hTEAD. For XIAP, the RISM-based analyses were successfully employed to explain SAR trends in a series of closely related ligands. In case of Bcl-xL, the same analyses were performed on a set of complex structures that were obtained by previously performed simulations of the protein with Ugi-type model compounds provided by the Brunschweiger group (done by J. Borchert). By analysing the water thermodynamics and the distribution of the pharmacophoric probes, suggestions for the design of respective libraries using the Ugi reaction were made which have to be validated by future experiments. For the highly demanding target hTEAD which exhibits several large binding grooves, the Brunschweiger group already obtained hit molecules in a screening that contain a Cl-indole moiety. By analysing available complex structures of peptides with respective SAR trends, a prediction of the most likely binding position of the hit molecule was made. Afterwards, docking of a set of model compounds for a click reaction-based screening library provided by the Brunschweiger group were docked and scored according to the developed workflow, and a proposal was made for the most promising building blocks. All in all, the analyses on the PPI targets showed that the approaches developed in this work together allow to get a complete picture of the thermodynamic signature of a protein binding site, thus allowing to generate design strategies for novel ligands or screening libraries.

In the future, the proposed strategies have to be validated by biochemical assays and ideally by co-crystallisation of respective model compounds. Besides, an iterative circle should be established, in which newly gained experimental insights are taken up to optimise the developed computational approach, which then in turn can be used for improved hit-to-lead development. Potentially, the integration of ML-based approaches could help to automatise such a workflow: Thanks to the possibility to interpolate 3D RISM-derived fields onto distinct ligand atoms, which was introduced in this work, respective atom-based descriptors are readily available and could be used with graph convolutional neural networks (CNN). Alternatively, whole fields could be used as input for 3D CNNs. Thus, ultimately, a scoring function could be trained to automatically guide the design of novel ligands for a given target protein structure.

W.r.t. water thermodynamics, it could be interesting to further investigate a local V_m- and q-based correction as proposed in 2.3.3.3. In this context, a direct comparison with alternative methods, for instance WaterMap, on the same protein structures would be helpful to further improve the calculation of absolute $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ contributions. In addition, it would also be of interest to achieve a further separation of the local information about $\Delta_{\mathrm{hyd}}G_{\mathrm{P}}$ into enthalpic and entropic contributions to gain further insights.

All in all, the approaches and workflows developed in this work were shown to be well-suited for driving ligand design and optimisation. They can complement the existing SBDD toolbox by adding valuable insight into effects that are not or only indirectly addressed by conventional methods.

6. Literature

1 Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular Determinants of Drug-Receptor Binding Kinetics. *Drug Discov. Today* **2013**, *18*. https://doi.org/10.1016/j.drudis.2013.02.007.

2 Burlingham, B. T.; Widlanski, T. S. An Intuitive Look at the Relationship of Ki and IC50: A More General Use for the Dixon Plot. *J. Chem. Educ.* **2003**, *80* (2), 214–218. https://doi.org/10.1021/ed080p214.

3 Olsson, T. S. G.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. The Thermodynamics of Protein-Ligand Interaction and Solvation: Insights for Ligand Design. *J. Mol. Biol.* **2008**, *384* (4), 1002–1017. https://doi.org/10.1016/j.jmb.2008.09.073.

4 Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72* (3), 1047–1069. https://doi.org/10.1016/S0006-3495(97)78756-3.

5 Dunitz, J. D. Win Some, Lose Some: Enthalpy-Entropy Compensation in Weak Intermolecular Interactions. *Chem. Biol.* **1995**, *2* (11), 709–712. https://doi.org/10.1016/1074-5521(95)90097-7.

6 Van Montfort, R. L. M.; Workman, P. Structure-Based Drug Design: Aiming for a Perfect Fit. *Essays Biochem.* **2017**, *61* (5), 431–437. https://doi.org/10.1042/EBC20170052.

7 Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. Berichte der Dtsch. Chem. Gesellschaft 1894, 27 (3), 2985–2993. https://doi.org/10.1002/cber.18940270364.

8 Batool, M.; Ahmad, B.; Choi, S. A Structure-Based Drug Discovery Paradigm. *Int. J. Mol. Sci.* **2019**, *20* (11), 2783. https://doi.org/10.3390/ijms20112783.

9 Perutz, M. F.; Rosa, J.; Schechter, A. Therapeutic Agents for Sickle Cell Disease. *Nature* **1978**, *275* (5679), 369–370. https://doi.org/10.1038/275369a0.

10 Thomas, S. E.; Mendes, V.; Kim, S. Y.; Malhotra, S.; Ochoa-Montaño, B.; Blaszczyk, M.; Blundell, T. L. Structural Biology and the Design of New Therapeutics: From HIV and Cancer to Mycobacterial Infections: A Paper Dedicated to John Kendrew. *J. Mol. Biol.* **2017**, *429* (17), 2677–2693. https://doi.org/10.1016/j.jmb.2017.06.014.

11 Jaskolski, M.; Dauter, Z.; Wlodawer, A. A Brief History of Macromolecular Crystallography, Illustrated by a Family Tree and Its Nobel Fruits. *FEBS J.* **2014**, *281* (18), 3985–4009. https://doi.org/10.1111/febs.12796.

12 Chen, D.; Jansson, A.; Sim, D.; Larsson, A.; Nordlund, P. Structural Analyses of Human Thymidylate Synthase Reveal a Site That May Control Conformational Switching between Active and Inactive States. J. Biol. Chem. 2017, 292 (32), 13449–13458. https://doi.org/10.1074/jbc.M117.787267.

13 King, N. M.; Prabu-Jeyabalan, M.; Bandaranayake, R. M.; Nalam, M. N. L.; Nalivaika, E. A.; Özen, A.; Haliloğlu, T.; Yilmaz, N. K.; Schiffer, C. A. Extreme Entropy-Enthalpy Compensation in a Drug-Resistant Variant of HIV-1 Protease. ACS Chem. Biol. 2012, 7 (9), 1536–1546. https://doi.org/10.1021/cb300191k.

14 Dessen, A.; Quémard, A.; Blanchard, J. S.; Jacobs, W. R.; Sacchettini, J. C. Crystal Structure and Function of the Isoniazid Target of Mycobacterium Tuberculosis. *Science (80-.).* **1995**, *267* (5204), 1638–1641. https://doi.org/10.1126/science.7886450.

15 Zhang, L.; Zhang, H.; Zhao, Y.; Li, Z.; Chen, S.; Zhai, J.; Chen, Y.; Xie, W.; Wang, Z.; Li, Q.; Zheng, X.; Hu, X. Inhibitor Selectivity between Aldo-Keto Reductase Superfamily Members AKR1B10 and AKR1B1: Role of Trp112 (Trp111). *FEBS Lett.* **2013**, *587* (22), 3681–3686. https://doi.org/10.1016/j.febslet.2013.09.031.

16 Kurumbail, R. G.; Stevens, A. M.; Gierse, J. K.; McDonald, J. J.; Stegeman, R. A.; Pak, J. Y.; Gildehaus, D.; Miyashiro, J. M.; Penning, T. D.; Seibert, K.; Isakson, P. C.; Stallings, W. C. Structural Basis for Selective Inhibition of Cyciooxygenase-2 by Anti-Inflammatory Agents. *Nature* **1996**, *384* (6610), 644–648. https://doi.org/10.1038/384644a0.

17 Ghersi, D.; Sanchez, R. Beyond Structural Genomics: Computational Approaches for the Identification of Ligand Binding Sites in Protein Structures. *J. Struct. Funct. Genomics* **2011**, *12* (2), 109–117.

https://doi.org/10.1007/s10969-011-9110-6.

18 Levitt, D. G.; Banaszak, L. J. POCKET: A Computer Graphies Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino Acids. *J. Mol. Graph.* **1992**, *10* (4), 229–234. https://doi.org/10.1016/0263-7855(92)80074-N.

19 Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins. *J. Mol. Graph. Model.* **1997**, *15* (6), 359–363. https://doi.org/10.1016/S1093-3263(98)00002-3.

20 Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graph.* **1995**, *13* (5), 323–330. https://doi.org/10.1016/0263-7855(95)00073-9.

21 Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849–857. https://doi.org/10.1021/jm00145a002.

22 Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: An Energy-Based Method for the Prediction of Protein-Ligand Binding Sites. *Bioinformatics* **2005**, *21* (9), 1908–1916. https://doi.org/10.1093/bioinformatics/bti315.

23 Morita, M.; Nakamura, S.; Shimizu, K. Highly Accurate Method for Ligand-Binding Site Prediction in Unbound State (Apo) Protein Structures. Proteins Struct. Funct. Bioinforma. 2008, 73 (2), 468–479. https://doi.org/10.1002/prot.22067.

24 Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1639**, *19* (14).

25 Harris, R.; Olson, A. J.; Goodsell, D. S. Automated Prediction of Ligand-Binding Sites in Proteins. *Proteins* **2007**, *70* (4), 1506–1517. https://doi.org/10.1002/prot.21645.

26 Böhm, H.-J. The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors. *J. Comput. Aided. Mol. Des.* **1992**, *6* (1), 61–78. https://doi.org/10.1007/BF00124387.

27 Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: A Program for Structure Generation. *J. Comput. Aided. Mol. Des.* **1993**, *7* (2), 127–153. https://doi.org/10.1007/BF00126441.

28 Mok, N. Y.; Chadwick, J.; Kellett, K. A. B.; Casas-Arce, E.; Hooper, N. M.; Johnson, A. P.; Fishwick, C. W. G. Discovery of Biphenylacetamide-Derived Inhibitors of BACE1 Using de Novo Structure-Based Molecular Design. *J. Med. Chem.* **2013**, *56* (5), 1843–1852. https://doi.org/10.1021/jm301127x.

29 Wang, R.; Gao, Y.; Lai, L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. J. Mol. Model. 2000, 6 (7–8), 498–516. https://doi.org/10.1007/s0089400060498.

30 Shang, E.; Yuan, Y.; Chen, X.; Liu, Y.; Pei, J.; Lai, L. De Novo Design of Multitarget Ligands with an Iterative Fragment-Growing Strategy. *J. Chem. Inf. Model.* **2014**, *54* (4), 1235–1241. https://doi.org/10.1021/ci500021v.

31 Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-Driven de Novo Design of Bioactive Compounds. *PLoS Comput. Biol.* **2012**, *8* (2), e1002380. https://doi.org/10.1371/journal.pcbi.1002380.

32 Friedrich, L.; Rodrigues, T.; Neuhaus, C. S.; Schneider, P.; Schneider, G. From Complex Natural Products to Simple Synthetic Mimetics by Computational De Novo Design. *Angew. Chemie Int. Ed.* **2016**, *55* (23), 6789–6792. https://doi.org/10.1002/anie.201601941.

33 Rodrigues, T.; Reker, D.; Welin, M.; Caldera, M.; Brunner, C.; Gabernet, G.; Schneider, P.; Walse, B.; Schneider, G. De-Novo-Fragmententwurf Für Die Wirkstoffforschung Und Chemische Biologie. *Angew. Chemie* **2015**, *127* (50), 15294–15298. https://doi.org/10.1002/ange.201508055.

34 Rodrigues, T.; Roudnicky, F.; Koch, C. P.; Kudoh, T.; Reker, D.; Detmar, M.; Schneider, G. De Novo Design and Optimization of Aurora A Kinase Inhibitors. *Chem. Sci.* **2013**, *4* (3), 1229–1233. https://doi.org/10.1039/c2sc21842a.

35 Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inform.* **2018**, *37* (1). https://doi.org/10.1002/minf.201700153.

36 Ehrlich, P. Über Den Jetzigen Stand Der Chemotherapie. *Berichte der Dtsch. Chem. Gesellschaft* **1909**,

42 (1), 17–47. https://doi.org/10.1002/cber.19090420105.

37 Kim, K.-H.; Kim, N. D.; Seong, B.-L. Pharmacophore-Based Virtual Screening: A Review of Recent Applications. *Expert Opin. Drug Discov.* **2010**, *5* (3), 205–222. https://doi.org/10.1517/17460441003592072.

38 Wermuth, C. G. Pharmacophores: Historical Perspective and Viewpoint from a Medicinal Chemist. In *Pharmacophores and Pharmacophore Searches*; Wiley, 2006; pp 1–13. https://doi.org/10.1002/3527609164.ch1.

39 Hou, T.; Wang, J.; Zhang, W.; Wang, W.; Xu, X. Recent Advances in Computational Prediction of Drug Absorption and Permeability in Drug Discovery. *Curr. Med. Chem.* **2006**, *13* (22), 2653–2667. https://doi.org/10.2174/092986706778201558.

40 Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. Predicting Molecular Interactions in Silico: I. A Guide to Pharmacophore Identification and Its Applications to Drug Design. *Curr. Med. Chem.* **2005**, *11* (1), 71–90. https://doi.org/10.2174/0929867043456287.

41 Sanders, M. P. A.; McGuire, R.; Roumen, L.; de Esch, I. J. P.; de Vlieg, J.; Klomp, J. P. G.; de Graaf, C. From the Protein’s Perspective: The Benefits and Challenges of Protein Structure-Based Pharmacophore Modeling. *Med. Chem. Commun.* **2012**, *3* (1), 28–38. https://doi.org/10.1039/C1MD00210D.

42 Schaller, D.; Šribar, D.; Noonan, T.; Deng, L.; Nguyen, T. N.; Pach, S.; Machalz, D.; Bermudez, M.; Wolber, G. Next Generation 3D Pharmacophore Modeling. *WIREs Comput. Mol. Sci.* **2020**, *10* (4). https://doi.org/10.1002/wcms.1468.

43 Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason‖, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model* **2007**, *47* (2), 279–294. https://doi.org/10.1021/CI600253E.

44 Chemical Computing Group. Molecular operating environment (MOE). Montreal, QC, Canada; 2010.

45 Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45* (1), 160–169. https://doi.org/10.1021/ci049885e.

46 Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J. Comput. Aided. Mol. Des.* **2006**, *20*, 647–671. https://doi.org/10.1007/s10822-006-9087-6.

47 Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations among Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (3), 563–571. https://doi.org/10.1021/ci950273r.

48 Tanimoto, T. T. An Elementary Mathematical Theory of Classification and Prediction. *Internal IBM Technical Report*. **1957**.

49 Sanders, M. P. A.; Barbosa, A. J. M.; Zarzycka, B.; Nicolaes, G. A. F.; Klomp, J. P. G.; De Vlieg, J.; Del Rio, A. Comparative Analysis of Pharmacophore Screening Tools. *J. Chem. Inf. Model.* **2012**, *52* (6), 1607–1620. https://doi.org/10.1021/ci2005274.

50 Huang, S.-Y.; Zou, X. Advances and Challenges in Protein-Ligand Docking. *Int. J. Mol. Sci.* **2010**, *11* (8), 3016–3034. https://doi.org/10.3390/ijms11083016.

51 Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

52 Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749. https://doi.org/10.1021/jm0306430.

53 Rarey, M.; Kramer, B.; Lengauer, T. Multiple Automatic Base Selection: Protein-Ligand Docking Based on Incremental Construction without Manual Intervention. *J. Comput. Aided. Mol. Des.* **1997**, *11* (4), 369–384. https://doi.org/10.1023/A:1007913026166.

54 Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288. https://doi.org/10.1016/0022-

2836(82)90153-X.

55 Goodsell, D. S.; Olson, A. J. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins* **1990**, *8* (3), 195–202. https://doi.org/10.1002/prot.340080302.

56 Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci.* **1958**, *44* (2), 98–104. https://doi.org/10.1073/pnas.44.2.98.

57 Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598–1603. https://doi.org/10.1126/science.1749933.

58 Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450*, 964–972. https://doi.org/10.1038/nature06522.

59 Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins* **1995**, *21* (3), 167–195. https://doi.org/10.1002/prot.340210302.

60 Miller, D. W.; Dill, K. A. Ligand Binding to Proteins: The Binding Landscape Model. *Protein Sci.* **1997**, *6* (10), 2166–2179. https://doi.org/10.1002/pro.5560061011.

61 Ma, B.; Kumar, S.; Tsai, C. J.; Nussinov, R. Folding Funnels and Binding Mechanisms. *Protein Eng.* **1999**, *12* (9), 713–720. https://doi.org/10.1093/protein/12.9.713.

62 Tsai, C.-J.; Kumar, S.; Ma, B.; Nussinov, R. Folding Funnels, Binding Funnels, and Protein Function. *Protein Sci.* **1999**, *8* (6), 1181–1190. https://doi.org/10.1110/ps.8.6.1181.

63 Tobi, D.; Bahar, I. Structural Changes Involved in Protein Binding Correlate with Intrinsic Motions of Proteins in the Unbound State. *PNAS* **2005**, *102* (52), 18908–18913. https://doi.org/10.1073/pnas.0507603102.

64 Csermely, P.; Palotai, R.; Nussinov, R. Induced Fit, Conformational Selection and Independent Dynamic Segments: An Extended View of Binding Events. *Trends Biochem. Sci.* **2010**, *35* (10), 539–546. https://doi.org/10.1016/j.tibs.2010.04.009.

65 Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The Penultimate Rotamer Library. *Proteins* **2000**, *40* (3), 389–408. https://doi.org/10.1002/1097-0134(20000815)40:3<389::AID-PROT50>3.0.CO;2-2.

66 Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. Four-Dimensional Docking: A Fast and Accurate Account of Discrete Receptor Flexibility in Ligand Docking. *J. Med. Chem.* **2009**, *52* (2), 397–406. https://doi.org/10.1021/jm8009958.

67 Broughton, H. B. A Method for Including Protein Flexibility in Protein-Ligand Docking: Improving Tools for Database Mining and Virtual Screening. *J. Mol. Graph. Model.* **2000**, *18* (3). https://doi.org/10.1016/S1093-3263(00)00036-X.

68 Lu, S. Y.; Jiang, Y. J.; Lv, J.; Zou, J. W.; Wu, T. X. Role of Bridging Water Molecules in GSK3β-Inhibitor Complexes: Insights from QM/MM, MD, and Molecular Docking Studies. *J. Comput. Chem.* **2011**, *32* (9), 1907–1918. https://doi.org/10.1002/jcc.21775.

69 Corbeil, C. R.; Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 1. Development and Validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, *47* (2), 435–449. https://doi.org/10.1021/ci6002637.

70 Carugo, O.; Bordo, D. How Many Water Molecules Can Be Detected by Protein Crystallography? *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1999**, *55* (2), 479–483. https://doi.org/10.1107/S0907444998012086.

71 Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50* (4), 726–741. https://doi.org/10.1021/jm061277y.

72 Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-based Energy Evaluation. *J. Comput. Chem.* **1992**, *13* (4), 505–524. https://doi.org/10.1002/jcc.540130412.

73 Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins* **2003**, *52* (4), 609–623. https://doi.org/10.1002/prot.10465.

74 Chen, F.; Liu, H.; Sun, H.; Pan, P.; Li, Y.; Li, D.; Hou, T. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 6. Capability to Predict Protein-Protein Binding Free Energies and Re-Rank

Binding Poses Generated by Protein-Protein Docking. *Phys. Chem. Chem. Phys.* **2016**, *18* (32), 22129–22139. https://doi.org/10.1039/c6cp03670h.

75 Yang, Y.; Lightstone, F. C.; Wong, S. E. Approaches to Efficiently Estimate Solvation and Explicit Water Energetics in Ligand Binding: The Use of WaterMap. *Expert Opinion on Drug Discovery*. Taylor & Francis March 2013, pp 277–287. https://doi.org/10.1517/17460441.2013.749853.

76 Ross, G. A.; Morris, G. M.; Biggin, P. C. Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites. *PLoS One* **2012**, *7* (3), e32036. https://doi.org/10.1371/journal.pone.0032036.

77 Uehara, S.; Tanaka, S. AutoDock-GIST: Incorporating Thermodynamics of Active-Site Water into Scoring Function for Accurate Protein-Ligand Docking. *Molecules* **2016**, *21* (11), 1604. https://doi.org/10.3390/molecules21111604.

78 Li, J.; Fu, A.; Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdiscip. Sci. Comput. Life Sci.* **2019**, *11* (2), 320–328. https://doi.org/10.1007/s12539-019-00327-w.

79 Chaskar, P.; Zoete, V.; Röhrig, U. F. On-the-Fly QM/MM Docking with Attracting Cavities. *J. Chem. Inf. Model.* **2017**, *57* (1), 73–84. https://doi.org/10.1021/acs.jcim.6b00406.

80 Chaskar, P.; Zoete, V.; Röhrig, U. F. Toward On-the-Fly Quantum Mechanical/Molecular Mechanical (QM/MM) Docking: Development and Benchmark of a Scoring Function. *J. Chem. Inf. Model.* **2014**, *54* (11), 3137–3152. https://doi.org/10.1021/ci5004152.

81 Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84–96. https://doi.org/10.1021/ci800298z.

82 Huang, S. Y.; Zou, X. Mean-Force Scoring Functions for Protein-Ligand Binding. In *Annual Reports in Computational Chemistry*; Elsevier BV, 2010; Vol. 6, pp 280–296. https://doi.org/10.1016/S1574-1400(10)06014-7.

83 Thomas, P. D.; Dill, K. A. An Iterative Method for Extracting Energy-like Quantities from Protein Structures. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93* (21), 11628–11633. https://doi.org/10.1073/pnas.93.21.11628.

84 Koppensteiner, W. A.; Sippl, M. J. Knowledge-Based Potentials - Back to the Roots. *Biochem.* **1998**, *63* (3), 247–252.

85 Thomas, P. D.; Dill, K. A. Statistical Potentials Extracted from Protein Structures: How Accurate Are They? *J. Mol. Biol.* **1996**, *257* (2), 457–469. https://doi.org/10.1006/jmbi.1996.0175.

86 Mooij, W. T. M.; Verdonk, M. L. General and Targeted Statistical Potentials for Protein-Ligand Interactions. *Proteins Struct. Funct. Genet.* **2005**, *61* (2), 272–287. https://doi.org/10.1002/prot.20588.

87 Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295* (2), 337–356. https://doi.org/10.1006/jmbi.1999.3371.

88 Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScoreCSD-Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of near-Native Ligand Poses and Better Affinity Prediction. *J. Med. Chem.* **2005**, *48* (20), 6296–6303. https://doi.org/10.1021/jm050436v.

89 Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42* (5), 791–804. https://doi.org/10.1021/jm980536j.

90 Muegge, I. PMF Scoring Revisited. *J. Med. Chem.* **2006**, *49* (20), 5895–5902. https://doi.org/10.1021/jm050038s.

91 Huang, S. Y.; Zou, X. An Iterative Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions: I. Derivation of Interaction Potentials. *J. Comput. Chem.* **2006**, *27* (15), 1866–1875. https://doi.org/10.1002/jcc.20504.

92 Huang, S. Y.; Zou, X. An Iterative Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions: II. Validation of the Scoring Function. *J. Comput. Chem.* **2006**, *27* (15), 1876–1882. https://doi.org/10.1002/jcc.20505.

93 Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A Knowledge-Based Energy Function for Protein-Ligand, Protein-Protein, and Protein-DNA Complexes. *J. Med. Chem.* **2005**, *48* (7), 2325–2335. https://doi.org/10.1021/jm049314d.

94 Huang, S. Y.; Zou, X. Inclusion of Solvation and Entropy in the Knowledge-Based Scoring Function for

Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2010**, *50* (2), 262–273. https://doi.org/10.1021/ci9002987.

95 Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26* (9), 1169–1175. https://doi.org/10.1093/bioinformatics/btq112.

96 Shen, C.; Ding, J.; Wang, Z.; Cao, D.; Ding, X.; Hou, T. From Machine Learning to Deep Learning: Advances in Scoring Functions for Protein–Ligand Docking. *WIREs Comput. Mol. Sci.* **2020**, *10* (1). https://doi.org/10.1002/wcms.1429.

97 Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a More Precise Chemical Description of Protein-Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* **2014**, *54* (3), 944–955. https://doi.org/10.1021/ci500091r.

98 Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today*. Elsevier Ltd August 1, 2018, pp 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010.

99 Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120–131. https://doi.org/10.1021/acscentsci.7b00512.

100 Karlov, D. S.; Sosnin, S.; Fedorov, M. V.; Popov, P. GraphDelta: MPNN Scoring Function for the Affinity Prediction of Protein-Ligand Complexes. *ACS Omega* **2020**, *5* (10), 5150–5159. https://doi.org/10.1021/acsomega.9b04162.

101 Ashtawy, H. M.; Mahapatra, N. R. A Comparative Assessment of Predictive Accuracies of Conventional and Machine Learning Scoring Functions for Protein-Ligand Binding Affinity Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2015**, *12* (2), 335–347. https://doi.org/10.1109/TCBB.2014.2351824.

102 Zilian, D.; Sotriffer, C. A. SFCscoreRF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53* (8), 1923–1933. https://doi.org/10.1021/ci400120b.

103 Li, H.; Leung, K. S.; Wong, M. H.; Ballester, P. J. Substituting Random Forest for Multiple Linear Regression Improves Binding Affinity Prediction of Scoring Functions: Cyscore as a Case Study. *BMC Bioinformatics* **2014**, *15* (1), 1–12. https://doi.org/10.1186/1471-2105-15-291.

104 Torng, W.; Altman, R. B. 3D Deep Convolutional Neural Networks for Amino Acid Environment Similarity Analysis. *BMC Bioinformatics* **2017**, *18* (1), 302. https://doi.org/10.1186/s12859-017-1702-0.

105 Shen, C.; Hu, Y.; Wang, Z.; Zhang, X.; Pang, J.; Wang, G.; Zhong, H.; Xu, L.; Cao, D.; Hou, T. Beware of the Generic Machine Learning-Based Scoring Functions in Structure-Based Virtual Screening. *Brief. Bioinform.* **2020**. https://doi.org/10.1093/bib/bbaa070.

106 Noble, W. S. What Is a Support Vector Machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567.

107 Li, G. B.; Yang, L. L.; Wang, W. J.; Li, L. L.; Yang, S. Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53* (3), 592–600. https://doi.org/10.1021/ci300493w.

108 Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inform.* **2015**, *34* (2–3), 115–126. https://doi.org/10.1002/minf.201400132.

109 Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38* (16), 1291–1307. https://doi.org/10.1002/jcc.24764.

110 Durrant, J. D.; McCammon, J. A. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2010**, *50* (10), 1865–1871. https://doi.org/10.1021/ci100244v.

111 Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor-Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51* (11), 2897–2903. https://doi.org/10.1021/ci2003889.

112 Durrant, J. D.; Carlson, K. E.; Martin, T. A.; Offutt, T. L.; Mayne, C. G.; Katzenellenbogen, J. A.; Amaro, R. E. Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor Ligands. *J.*

Chem. Inf. Model. **2015**, *55* (9), 1953–1961. https://doi.org/10.1021/acs.jcim.5b00241.

113 Lindert, S.; Zhu, W.; Liu, Y.; Pang, R.; Oldfield, E.; McCammon, J. A. Farnesyl Diphosphate Synthase Inhibitors from *In Silico* Screening. *Chem. Biol. Drug Des.* **2013**, *81* (6), 742–748. https://doi.org/10.1111/cbdd.12121.

114 Durrant, J. D.; Amaro, R. E. Machine-Learning Techniques Applied to Antibacterial Drug Discovery. *Chem. Biol. Drug Des.* **2015**, *85* (1), 14–21. https://doi.org/10.1111/cbdd.12423.

115 Skaff, D. A.; McWhorter, W. J.; Geisbrecht, B. V.; Wyckoff, G. J.; Miziorko, H. M. Inhibition of Bacterial Mevalonate Diphosphate Decarboxylase by Eriochrome Compounds. *Arch. Biochem. Biophys.* **2015**, *566*, 1–6. https://doi.org/10.1016/j.abb.2014.12.002..

116 Buryska, T.; Daniel, L.; Kunka, A.; Brezovsky, J.; Damborsky, J.; Prokop, Z. Discovery of Novel Haloalkane Dehalogenase Inhibitors. *Appl. Environ. Microbiol.* **2016**, *82* (6), 1958–1965. https://doi.org/10.1128/AEM.03916-15.

117 Albawi, S.; Mohammed, T. A.; Al-Zawi, S. Understanding of a Convolutional Neural Network. In *Proceedings of 2017 International Conference on Engineering and Technology, ICET 2017*; Institute of Electrical and Electronics Engineers Inc., 2017. https://doi.org/10.1109/ICEngTechnol.2017.8308186.

118 Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. *arXiv* **2015**.

119 Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58* (2), 287–296. https://doi.org/10.1021/acs.jcim.7b00650.

120 Gabel, J.; Desaphy, J.; Rognan, D. Beware of Machine Learning-Based Scoring Functions-on the Danger of Developing Black Boxes. *J. Chem. Inf. Model.* **2014**, *54* (10), 2807–2815. https://doi.org/10.1021/ci500406k.

121 Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47* (12), 2977–2980. https://doi.org/10.1021/jm030580l.

122 Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49* (4), 1079–1093. https://doi.org/10.1021/ci9000053.

123 Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, *54* (6), 1700–1716. https://doi.org/10.1021/ci500080q.

124 Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* **2015**, *31* (3), 405–412. https://doi.org/10.1093/bioinformatics/btu626.

125 Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, *59* (2), 895–913. https://doi.org/10.1021/acs.jcim.8b00545.

126 Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801. https://doi.org/10.1021/jm0608356.

127 Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. https://doi.org/10.1021/jm300687e.

128 Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49* (2), 169–184. https://doi.org/10.1021/ci8002649.

129 Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 - A Public Library of Challenging Docking Benchmark Sets. *J. Chem. Inf. Model.* **2013**, *53* (6), 1447–1462. https://doi.org/10.1021/ci400115b.

130 Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. Discovery of a Novel Binding Trench in HIV Integrase. *J. Med. Chem.* **2004**, *47* (8), 1879–1881.

https://doi.org/10.1021/jm0341913.

131 Hazuda, D. J.; Anthony, N. J.; Gomez, R. P.; Jolly, S. M.; Wai, J. S.; Zhuang, L.; Fisher, T. E.; Embrey, M.; Guare, J. P.; Egbertson, M. S.; Vacca, J. P.; Huff, J. R.; Felock, P. J.; Witmer, M. V.; Stillmock, K. A.; Danovich, R.; Grobler, J.; Miller, M. D.; Espeseth, A. S.; Jin, L.; Chen, I. W.; Lin, J. H.; Kassahun, K.; Ellis, J. D.; Wong, B. K.; Xu, W.; Pearson, P. G.; Schleif, W. A.; Cortese, R.; Emini, E.; Summa, V.; Holloway, M. K.; Young, S. D. A Naphthyridine Carboxamide Provides Evidence for Discordant Resistance between Mechanistically Identical Inhibitors of HIV-1 Integrase. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (31), 11233–11238. https://doi.org/10.1073/pnas.0402357101.

132 Xu, M.; Unzue, A.; Dong, J.; Spiliotopoulos, D.; Nevado, C.; Caflisch, A. Discovery of CREBBP Bromodomain Inhibitors by High-Throughput Docking and Hit Optimization Guided by Molecular Dynamics. *J. Med. Chem.* **2016**, *59* (4), 1340–1349. https://doi.org/10.1021/acs.jmedchem.5b00171.

133 Huang, D.; Caflisch, A. The Free Energy Landscape of Small Molecule Unbinding. *PLoS Comput. Biol.* **2011**, *7* (2), e1002002. https://doi.org/10.1371/journal.pcbi.1002002.

134 Huang, D.; Caflisch, A. Small Molecule Binding to Proteins: Affinity and Binding/Unbinding Dynamics from Atomistic Simulations. *ChemMedChem* **2011**, *6* (9), 1578–1580. https://doi.org/10.1002/cmdc.201100237.

135 Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *PNAS* **2011**, *108* (25), 10184–10189. https://doi.org/10.1073/pnas.1103547108.

136 Lewars, E. G. Molecular Mechanics. In *Computational Chemistry*; Springer International Publishing, 2016; pp 51–99. https://doi.org/10.1007/978-3-319-30916-3_3.

137 Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217. https://doi.org/10.1002/jcc.540040211.

138 Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614. https://doi.org/10.1002/jcc.21287.

139 Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

140 Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

141 Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676. https://doi.org/10.1002/jcc.20090.

142 Jorgensen, W. L.; Tirado-Rives, J. The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666. https://doi.org/10.1021/ja00214a001.

143 Halgren, T. A. MMFF VI. MMFF94s Option for Energy Minimization Studies. *J. Comput. Chem.* **1999**, *20* (7), 720–729. https://doi.org/10.1002/(SICI)1096-987X(199905)20:7<720::AID-JCC7>3.0.CO;2-X.

144 Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

145 van Gunsteren, W. F.; Berendsen, H. J. C. Moleküldynamik-Computersimulationen; Methodik, Anwendungen Und Perspektiven in Der Chemie. *Angew. Chemie* **1990**, *102* (9), 1020–1055. https://doi.org/10.1002/ange.19901020907.

146 Verlet, L. Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones

Molecules. *Phys. Rev.* **1967**, *159* (1), 98–103. https://doi.org/10.1103/PhysRev.159.98.

147 Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-Rich Systems. *J. Comput. Chem.* **1999**, *20* (8).

148 Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron*. Cell Press September 19, 2018, pp 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011.

149 Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695–1697. https://doi.org/10.1103/PhysRevA.31.1695.

150 Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690. https://doi.org/10.1063/1.448118.

151 Wu, X.; Brooks, B. R.; Vanden-Eijnden, E. Self-Guided Langevin Dynamics via Generalized Langevin Equation. *J. Comput. Chem.* **2016**, *37* (6), 595–601. https://doi.org/10.1002/jcc.24015.

152 Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

153 Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72* (4), 2384–2393. https://doi.org/10.1063/1.439486.

154 De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035–4061. https://doi.org/10.1021/acs.jmedchem.5b01684.

155 Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. The Relaxed Complex Method: Accommodating Receptor Flexibility for Drug Design with an Improved Scoring Scheme. *Biopolymers* **2003**, *68* (1), 47–62. https://doi.org/10.1002/bip.10218.

156 Lu, H.; Tonge, P. J. Drug-Target Residence Time: Critical Information for Lead Optimization. *Curr. Opin. Chem. Biol.* **2010**, *14* (4), 467–474. https://doi.org/10.1016/j.cbpa.2010.06.176.

157 Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *Cit. J. Chem. Phys.* **1985**, *83*, 3050. https://doi.org/10.1063/1.449208.

158 Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

159 Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. Chem. Phys. Lett. 1999, 314, 141–151.

160 Isralewitz, B.; Gao, M.; Schulten, K. Steered Molecular Dynamics and Mechanical Functions of Proteins. *Curr. Opin. Struct. Biol.* **2001**, *11* (2), 224–230. https://doi.org/10.1016/S0959-440X(00)00194-9.

161 Kollman, P. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev* **1993**, *93*, 2395–2417.

162 Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chemical Reviews*. American Chemical Society 2015, pp 6312–6356. https://doi.org/10.1021/cr5000283.

163 Ben-Naim, A. *Molecular Theory of Solutions*; Oxford University Press: New York, **2006**.

164 Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

165 Price, D. J.; Brooks, C. L. A Modified TIP3P Water Potential for Simulation with Ewald Summation. *J. Chem. Phys.* **2004**, *121* (20), 10096–10103. https://doi.org/10.1063/1.1808117.

166 Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction Models for Water in Relation to Protein Hydration. In *Intermolecular Forces*; Springer, Dordrecht, 1981; pp 331–342. https://doi.org/10.1007/978-94-015-7658-1_21.

167 Iftimie, R.; Minary, P.; Tuckerman, M. E. Ab Initio Molecular Dynamics: Concepts, Recent Developments, and Future Trends. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (19), 6654–6659.

https://doi.org/10.1073/pnas.0500193102.

168 Kleinjung, J.; Fraternali, F. Design and Application of Implicit Solvent Models in Biomolecular Simulations. *Curr. Opin. Struct. Biol.* **2014**, *25*, 126–134. https://doi.org/10.1016/j.sbi.2014.04.003.

169 Roux, B.; Simonson, T. Implicit Solvent Models. *Biophys. Chem.* **1999**, *78* (1–2), 1–20. https://doi.org/10.1016/S0301-4622(98)00226-9.

170 Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews*. American Chemical Society August 28, 2019, pp 9478–9508. https://doi.org/10.1021/acs.chemrev.9b00055.

171 Fogolari, F.; Brigo, A.; Molinari, H. The Poisson-Boltzmann Equation for Biomolecular Electrostatics: A Tool for Structural Biology. *J. Mol. Recognit.* **2002**, *15* (6), 377–392. https://doi.org/10.1002/jmr.577.

172 Bashford, D.; Case, D. A. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152. https://doi.org/10.1146/annurev.physchem.51.1.129.

173 Onufriev, A. Chapter 7 Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview. In *Annual Reports in Computational Chemistry*; Elsevier BV, 2008; Vol. 4, pp 125–137. https://doi.org/10.1016/S1574-1400(08)00007-8.

174 Ferrara, P.; Apostolakis, J.; Caflisch, A. Evaluation of a Fast Implicit Solvent Model for Molecular Dynamics Simulations. *Proteins* **2002**, *46* (1), 24–33. https://doi.org/10.1002/prot.10001.

175 Lazaridis, T. Implicit Solvent Simulations of Peptide Interactions with Anionic Lipid Membranes. *Proteins* **2004**, *58* (3), 518–527. https://doi.org/10.1002/prot.20358.

176 Feig, M.; Brooks, C. L. Recent Advances in the Development and Application of Implicit Solvent Models in Biomolecule Simulations. *Curr. Opin. Struct. Biol.* **2004**, *14* (2), 217–224. https://doi.org/10.1016/j.sbi.2004.03.009.

177 Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864. https://doi.org/10.1103/PhysRev.136.B864.

178 Evans, R. The Nature of the Liquid-Vapour Interface and Other Topics in the Statistical Mechanics of Non-Uniform, Classical Fluids. *Adv. Phys.* **1979**, *28* (2), 143–200. https://doi.org/10.1080/00018737900101365.

179 Hansen, J.-P.; McDonald, I. R. *Theory of Simple Liquids - 3rd Edition*; Elsevier, **2007**.

180 Wu, J. Classical Density Functional Theory for Molecular Systems. In *Variational Methods in Molecular Modeling*; Springer, Singapore, 2017; pp 65–99. https://doi.org/10.1007/978-981-10-2502-0_3.

181 Sauer, E.; Gross, J. Classical Density Functional Theory for Liquid-Fluid Interfaces and Confined Systems: A Functional for the Perturbed-Chain Polar Statistical Associating Fluid Theory Equation of State. *Ind. Eng. Chem. Res.* **2017**, *56* (14), 4119–4135. https://doi.org/10.1021/acs.iecr.6b04551.

182 Monson, P. A.; Morriss, G. P. Recent Progress in the Statistical Mechanics of Interaction Site Fluids. *Adv. Chem. Phys.* **1990**, *77*, 451–550.

183 Hirata, F. *Molecular Theory of Solvation*; Kluwer Academic Publishers, **2003**. https://doi.org/10.1007/1-4020-2590-4_1.

184 Chandler, D.; Andersen, H. C. Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids. *J. Chem. Phys.* **1972**, *57* (5), 1930–1937. https://doi.org/10.1063/1.1678513.

185 Hirata, F.; Rossky, P. J.; Montgomery Pettitt, B. The Interionic Potential of Mean Force in a Molecular Polar Solvent from an Extended RISM Equation. *J. Chem. Phys.* **1983**, *78* (6), 4133–4144. https://doi.org/10.1063/1.445090.

186 Beglov, D.; Roux, B. Solvation of Complex Molecules in a Polar Liquid: An Integral Equation Theory. *J. Chem. Phys.* **1996**, *104* (21), 8678–8689. https://doi.org/10.1063/1.471557.

187 Kovalenko, A.; Hirata, F. Three-Dimensional Density Profiles of Water in Contact with a Solute of Arbitrary Shape: A RISM Approach. *Chem. Phys. Lett.* **1998**, *290* (1–3), 237–244. https://doi.org/10.1016/S0009-2614(98)00471-0.

188 Kast, S. M.; Kloss, T. Closed-Form Expressions of the Chemical Potential for Integral Equation Closures with Certain Bridge Functions. *J. Chem. Phys.* **2008**, *129* (23), 236101.

https://doi.org/10.1063/1.3041709.

189 Joung, I. S.; Luchko, T.; Case, D. A. Simple Electrolyte Solutions: Comparison of DRISM and Molecular Dynamics Results for Alkali Halide Solutions. *J. Chem. Phys.* **2013**, *138* (4), 44103. https://doi.org/10.1063/1.4775743.

190 Kast, S. M. Free Energies from Integral Equation Theories: Enforcing Path Independence. *Phys. Rev. E* **2003**, *67* (4), 041203. https://doi.org/10.1103/PhysRevE.67.041203.

191 Ben-Naim, A.; Marcus, Y. Solvation Thermodynamics of Nonionic Solutes. *J. Chem. Phys.* **1984**, *81* (4), 2016–2027. https://doi.org/10.1063/1.447824.

192 Güssregen, S.; Matter, H.; Hessler, G.; Lionta, E.; Heil, J.; Kast, S. M. Thermodynamic Characterization of Hydration Sites from Integral Equation-Derived Free Energy Densities: Application to Protein Binding Sites and Ligand Series. *J. Chem. Inf. Model.* **2017**, *57* (7), 1652–1666. https://doi.org/10.1021/acs.jcim.6b00765.

193 Kloss, T.; Heil, J.; Kast, S. M. Quantum Chemistry in Solution by Combining 3D Integral Equation Theory with a Cluster Embedding Approach. *J. Phys. Chem. B* **2008**, *112* (14), 4337–4343. https://doi.org/10.1021/jp710680m.

194 Pongratz, T.; Kibies, P.; Eberlein, L.; Tielker, N.; Hölzl, C.; Imoto, S.; Beck Erlach, M.; Kurrmann, S.; Schummel, P. H.; Hofmann, M.; Reiser, O.; Winter, R.; Kremer, W.; Kalbitzer, H. R.; Marx, D.; Horinek, D.; Kast, S. M. Pressure-Dependent Electronic Structure Calculations Using Integral Equation-Based Solvation Models. *Biophys. Chem.* **2020**, *257*, 106258. https://doi.org/10.1016/j.bpc.2019.106258.

195 Tielker, N.; Tomazic, D.; Heil, J.; Kloss, T.; Ehrhart, S.; Güssregen, S.; Schmidt, K. F.; Kast, S. M. The SAMPL5 Challenge for Embedded-Cluster Integral Equation Theory: Solvation Free Energies, Aqueous PK a, and Cyclohexane–Water Log D. *J. Comput. Aided. Mol. Des.* **2016**, *30* (11), 1035–1044. https://doi.org/10.1007/s10822-016-9939-7.

196 Sergiievskyi, V.; Jeanmairet, G.; Levesque, M.; Borgis, D. Solvation Free-Energy Pressure Corrections in the Three Dimensional Reference Interaction Site Model. *J. Chem. Phys.* **2015**, *143* (18), 184116. https://doi.org/10.1063/1.4935065.

197 Misin, M.; Fedorov, M. V.; Palmer, D. S. Hydration Free Energies of Molecular Ions from Theory and Simulation. *J. Phys. Chem. B* **2016**, *120* (5), 975–983. https://doi.org/10.1021/acs.jpcb.5b10809.

198 Tielker, N.; Tomazic, D.; Eberlein, L.; Güssregen, S.; Kast, S. M. The SAMPL6 Challenge on Predicting Octanol–Water Partition Coefficients from EC-RISM Theory. *J. Comput. Aided. Mol. Des.* **2020**, *34* (4), 453–461. https://doi.org/10.1007/s10822-020-00283-4.

199 Imai, T. Molecular Theory of Partial Molar Volume and Its Applications to Biomolecular Systems. *Condens. Matter Phys.* **2007**, *3* (51), 343–361.

200 Ladbury, J. E. Just Add Water! The Effect of Water on the Specificity of Protein- Ligand Binding Sites and Its Potential Application to Drug Design. *Chemistry and Biology*. Elsevier Ltd December 1, 1996, pp 973–980. https://doi.org/10.1016/S1074-5521(96)90164-7.

201 Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C. H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. Rational Design of Potent, Bioavailable, Nonpeptide Cyclic Ureas as HIV Protease Inhibitors. *Science.* **1994**, *263* (5145), 380–384. https://doi.org/10.1126/science.8278812.

202 Liu, C.; Wrobleski, S. T.; Lin, J.; Ahmed, G.; Metzger, A.; Wityak, J.; Gillooly, K. M.; Shuster, D. J.; McIntyre, K. W.; Pitt, S.; Shen, D. R.; Zhang, R. F.; Zhang, H.; Doweyko, A. M.; Diller, D.; Henderson, I.; Barrish, J. C.; Dodd, J. H.; Schieven, G. L.; Leftheris, K. 5-Cyanopyrimidine Derivatives as a Novel Class of Potent, Selective, and Orally Active Inhibitors of P38a MAP Kinase. *J. Med. Chem.* **2005**, *48* (20), 6261–6270. https://doi.org/10.1021/jm0503594.

203 Lu, Y.; Wang, R.; Yang, C. Y.; Wang, S. Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structures of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2007**, *47* (2), 668–675. https://doi.org/10.1021/ci6003527.

204 García-Sosa, A. T. Hydration Properties of Ligands and Drugs in Protein Binding Sites: Tightly-Bound, Bridging Water Molecules and Their Effects and Consequences on Molecular Design Strategies. *J. Chem. Inf. Model.* **2013**, *53* (6), 1388–1405. https://doi.org/10.1021/ci3005786.

205 Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Energetics of Displacing Water Molecules from Protein

Binding Sites: Consequences for Ligand Optimization. *J. Am. Chem. Soc.* **2009**, *131* (42), 15403–15411. https://doi.org/10.1021/ja906058w.

206 Krimmer, S. G.; Cramer, J.; Betz, M.; Fridh, V.; Karlsson, R.; Heine, A.; Klebe, G. Rational Design of Thermodynamic and Kinetic Binding Profiles by Optimizing Surface Water Networks Coating Protein-Bound Ligands. *J. Med. Chem.* **2016**, *59* (23), 10530–10548. https://doi.org/10.1021/acs.jmedchem.6b00998.

207 Snyder, P. W.; Mecinović, J.; Moustakas, D. T.; Thomas, S. W.; Harder, M.; Mack, E. T.; Lockett, M. R.; Héroux, A.; Sherman, W.; Whitesides, G. M. Mechanism of the Hydrophobic Effect in the Biomolecular Recognition of Arylsulfonamides by Carbonic Anhydrase. *PNAS* **2011**, *108* (44), 17889–17894. https://doi.org/10.1073/pnas.1114107108.

208 Spyrakis, F.; Ahmed, M. H.; Bayden, A. S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**, *60* (16), 6781–6828. https://doi.org/10.1021/acs.jmedchem.7b00057.

209 Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein-Ligand Binding. *PNAS* **2007**, *104* (3), 808–813. https://doi.org/10.1073/pnas.0610202104.

210 Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, *130* (9), 2817–2831. https://doi.org/10.1021/ja0771033.

211 Abel, R.; Salam, N. K.; Shelley, J.; Farid, R.; Friesner, R. A.; Sherman, W. Contribution of Explicit Solvent Effects to the Binding Affinity of Small-Molecule Inhibitors in Blood Coagulation Factor Serine Proteases. *ChemMedChem* **2011**, *6*, 1049–1066. https://doi.org/10.1002/cmdc.201000533.

212 Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. Predicting Conserved Water-Mediated and Polar Ligand Interactions in Proteins Using a k-Nearest-Neighbors Genetic Algorithm. *J. Mol. Biol.* **1997**, *265* (4), 445–464. https://doi.org/10.1006/jmbi.1996.0746.

213 García-Sosa, A. T.; Mancera, R. L.; Dean, P. M. WaterScore: A Novel Method for Distinguishing between Bound and Displaceable Water Molecules in the Crystal Structure of the Binding Site of Protein-Ligand Complexes. *J. Mol. Model.* **2003**, *9* (3), 172–182. https://doi.org/10.1007/s00894-003-0129-x.

214 Patel, H.; Grüning, B. A.; Günther, S.; Merfort, I. PyWATER: A PyMOL Plug-in to Find Conserved Water Molecules in Proteins by Clustering. *Bioinformatics* **2014**, *30* (20), 2978–2980. https://doi.org/10.1093/bioinformatics/btu424.

215 Sanschagrin, P. C.; Kuhn, L. A. Cluster Analysis of Consensus Water Sites in Thrombin and Trypsin Shows Conservation between Serine Proteases and Contributions to Ligand Specificity. *Protein Sci.* **1998**, *7* (10), 2054–2064. https://doi.org/10.1002/pro.5560071002.

216 Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. Robust Classification of “Relevant” Water Molecules in Putative Protein Binding Sites. *J. Med. Chem.* **2008**, *51* (4), 1063–1067. https://doi.org/10.1021/jm701023h.

217 Kellogg, G. E.; Fornabaio, M.; Chen, D. L.; Abraham, D. J.; Kellogg, G. E.; Fornabaio, M.; Chen, D. L.; Abraham, D. J. New Application Design for a 3D Hydropathic Map-Based Search for Potential Water Molecules Bridging between Protein and Ligand. *Internet Electron. J. Mol. Des.* **2005**, *4* (3), 194–209.

218 Pitt, W. R.; Goodfellow, J. M. Modelling of Solvent Positions around Polar Groups in Proteins. *Protein Eng. Des. Sel.* **1991**, *4* (5), 531–537. https://doi.org/10.1093/protein/4.5.531.

219 Rossato, G.; Ernst, B.; Vedani, A.; Smieško, M. AcquaAlta: A Directional Approach to the Solvation of Ligand-Protein Complexes. *J. Chem. Inf. Model.* **2011**, *51* (8), 1867–1881. https://doi.org/10.1021/ci200150p.

220 Xiao, W.; He, Z.; Sun, M.; Li, S.; Li, H. Statistical Analysis, Investigation, and Prediction of the Water Positions in the Binding Sites of Proteins. *J. Chem. Inf. Model.* **2017**, *57* (7), 1517–1528. https://doi.org/10.1021/acs.jcim.6b00620.

221 Nittinger, E.; Flachsenberg, F.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. *J. Chem. Inf. Model.* **2018**,

58 (8), 1625–1637. https://doi.org/10.1021/acs.jcim.8b00271.

222 Bui, H.-H.; Schiewe, A. J.; Haworth, I. S. WATGEN: An Algorithm for Modeling Water Networks at Protein-Protein Interfaces. *J. Comput. Chem.* **2007**, *28* (14), 2241–2251. https://doi.org/10.1002/jcc.20751.

223 Miao, Y.; Baudry, J. Active-Site Hydration and Water Diffusion in Cytochrome P450cam: A Highly Dynamic Process. *Biophys. J.* **2011**, *101* (6), 1493–1503. https://doi.org/10.1016/j.bpj.2011.08.020.

224 de Beer, S.; Vermeulen, N.; Oostenbrink, C. The Role of Water Molecules in Computational Drug Design. *Curr. Top. Med. Chem.* **2010**, *10* (1), 55–66. https://doi.org/10.2174/156802610790232288.

225 Bernardi, R. C.; Melo, M. C. R.; Schulten, K. Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochim. Biophys. Acta - Gen. Subj.* **2015**, *1850* (5), 872–877. https://doi.org/10.1016/j.bbagen.2014.10.019.

226 Lazaridis, T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. *J. Phys. Chem. B* **1998**, *102* (18), 3531–3541. https://doi.org/10.1021/jp9723574.

227 Lazaridis, T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids. *J. Phys. Chem. B* **1998**, *102* (18), 3542–3550. https://doi.org/10.1021/jp972358w.

228 Nguyen, C. N.; Cruz, A.; Gilson, M. K.; Kurtzman, T. Thermodynamics of Water in an Enzyme Active Site: Grid-Based Hydration Analysis of Coagulation Factor Xa. *J. Chem. Theory Comput.* **2014**, *10* (7), 2769–2780. https://doi.org/10.1021/ct401110x.

229 Matter, H.; Güssregen, S. Characterizing Hydration Sites in Protein-Ligand Complexes towards the Design of Novel Ligands. *Bioorganic Med. Chem. Lett.* **2018**, *28* (14), 2343–2352. https://doi.org/10.1016/j.bmcl.2018.05.061.

230 Nguyen, C. N.; Kurtzman, T.; Gilson, M. K. Spatial Decomposition of Translational Water-Water Correlation Entropy in Binding Pockets. *J. Chem. Theory Comput.* **2016**, *12* (1), 414–429. https://doi.org/10.1021/acs.jctc.5b00939.

231 Nguyen, C. N.; Kurtzman Young, T.; Gilson, M. K. Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor Cucurbit[7]Uril. *J. Chem. Phys.* **2012**, *137* (4), 973–980. https://doi.org/10.1063/1.4733951.

232 Huggins, D. J.; Payne, M. C. Assessing the Accuracy of Inhomogeneous Fluid Solvation Theory in Predicting Hydration Free Energies of Simple Solutes. *J. Phys. Chem. B* **2013**, *117* (27), 8232–8244. https://doi.org/10.1021/jp4042233.

233 Schauperl, M.; Podewitz, M.; Waldner, B. J.; Liedl, K. R. Enthalpic and Entropic Contributions to Hydrophobicity. *J. Chem. Theory Comput.* **2016**, *12* (9), 4600–4610. https://doi.org/10.1021/acs.jctc.6b00422.

234 Li, Z.; Lazaridis, T. Computing the Thermodynamic Contributions of Interfacial Water. In *Computational Drug Discovery and Design. Methods in Molecular Biology (Methods and Protocols)*; Baron, R., Ed.; Springer, New York, NY: New York, 2012; Vol. 819, pp 393–404. https://doi.org/10.1007/978-1-61779-465-0_24.

235 Cui, G.; Swails, J. M.; Manas, E. S. SPAM: A Simple Approach for Profiling Bound Water Molecules. *J. Chem. Theory Comput.* **2013**, *9* (12), 5539–5549. https://doi.org/10.1021/ct400711g.

236 Hu, B.; Lill, M. A. WATsite: Hydration Site Prediction Program with PyMOL Interface. *J. Comput. Chem.* **2014**, *35* (16), 1255–1260. https://doi.org/10.1002/jcc.23616.

237 Lin, S. T.; Maiti, P. K.; Goddard, W. A. Two-Phase Thermodynamic Model for Efficient and Accurate Absolute Entropy of Water from Molecular Dynamics Simulations. *J. Phys. Chem. B* **2010**, *114* (24), 8191–8198. https://doi.org/10.1021/jp103120q.

238 Henchman, R. H. Free Energy of Liquid Water from a Computer Simulation via Cell Theory. *J. Chem. Phys.* **2007**, *126* (6). https://doi.org/10.1063/1.2434964.

239 Adams, D. J. Chemical Potential of Hard-Sphere Fluids by Monte Carlo Methods. *Mol. Phys.* **1974**, *28* (5), 1241–1252. https://doi.org/10.1080/00268977400102551.

240 Adams, D. J. Grand Canonical Ensemble Monte Carlo for a Lennard-Jones Fluid. *Mol. Phys.* **1975**, *29* (1), 307–311. https://doi.org/10.1080/00268977500100221.

241 Guarnieri, F.; Mezei, M. Simulated Annealing of Chemical Potential: A General Procedure for Leading

Bound Waters. Application to the Study of the Differential Hydration Propensities of the Major and Minor Grooves of DNA. *J. Am. Chem. Soc.* **1996**, *118* (35), 8493–8494. https://doi.org/10.1021/ja961482a.

242 Rakhmanov, S. V.; Makeev, V. J. Atomic Hydration Potentials Using a Monte Carlo Reference State (MCRS) for Protein Solvation Modeling. *BMC Struct. Biol.* **2007**, *7* (1), 1–17. https://doi.org/10.1186/1472-6807-7-19.

243 Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Prediction of the Water Content in Protein Binding Sites. *J. Phys. Chem. B* **2009**, *113* (40), 13337–13346. https://doi.org/10.1021/jp9047456.

244 Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **2007**, *129* (9), 2577–2587. https://doi.org/10.1021/ja066980q.

245 SZMAP, 1.0.0; OpenEye Scientific Software Inc.: Santa Fe, NM, USA, **2011**.

246 Bayden, A. S.; Moustakas, D. T.; Joseph-McCarthy, D.; Lamb, M. L. Evaluating Free Energies of Binding and Conservation of Crystallographic Waters Using SZMAP. *J. Chem. Inf. Model.* **2015**, *55* (8), 1552–1565. https://doi.org/10.1021/ci500746d.

247 Chandler, D.; Andersen, H. C. Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids. *J. Chem. Phys.* **1972**, *57* (5), 1918–1929. https://doi.org/10.1063/1.1678512.

248 Yamazaki, T.; Kovalenko, A. Spatial Decomposition of Solvation Free Energy Based on the 3D Integral Equation Theory of Molecular Liquid: Application to Miniproteins. *J. Phys. Chem. B* **2011**, *115* (2), 310–318. https://doi.org/10.1021/jp1082938.

249 Huang, W.; Dedzo, G. K.; Stoyanov, S. R.; Lyubimova, O.; Gusarov, S.; Singh, S.; Lao, H.; Kovalenko, A.; Detellier, C. Molecule-Surface Recognition between Heterocyclic Aromatic Compounds and Kaolinite in Toluene Investigated by Molecular Theory of Solvation and Thermodynamic and Kinetic Experiments. *J. Phys. Chem. C* **2014**, *118* (41), 23821–23834. https://doi.org/10.1021/jp507393u.

250 Sindhikara, D. J.; Hirata, F. Analysis of Biomolecular Solvation Sites by 3D-RISM Theory. *J. Phys. Chem. B* **2013**, *117* (22), 6718–6723. https://doi.org/10.1021/jp4046116.

251 Kiyota, Y.; Takeda-Shitaka, M. Molecular Recognition Study on the Binding of Calcium to Calbindin D9k Based on 3D Reference Interaction Site Model Theory. *J. Phys. Chem. B* **2014**, *118* (39), 11496–11503. https://doi.org/10.1021/jp504822r.

252 O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (10), 1–14. https://doi.org/10.1186/1758-2946-3-33.

253 D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman (**2018**), AMBER 2018, University of California, San Francisco.

254 Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21* (2), 132–146. https://doi.org/10.1002/(SICI)1096-987X(20000130)21:2<132::AID-JCC5>3.0.CO;2-P.

255 Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641. https://doi.org/10.1002/jcc.10128.

256 Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*; Kluwer Academic Publishers, 2008; pp 319–326. https://doi.org/10.1007/978-3-540-78246-9_38.

257 Maw, S.; Sato, H.; Ten-no, S.; Hirata, F. Ab Initio Study of Water: Self-Consistent Determination of Electronic Structure and Liquid State Properties. *Chem. Phys. Lett.* **1997**, *276* (1–2), 20–25. https://doi.org/10.1016/S0009-2614(97)88029-3.

258 Kast, S. M.; Heil, J.; Güssregen, S.; Schmidt, K. F. Prediction of Tautomer Ratios by Embedded-Cluster Integral Equation Theory. *J. Comput. Aided. Mol. Des.* **2010**, *24* (4), 343–353.

https://doi.org/10.1007/s10822-010-9340-x.

259 Mrugalla, F.; Kast, S. M. Designing Molecular Complexes Using Free-Energy Derivatives from Liquid-State Integral Equation Theory. *J. Phys. Condens. Matter* **2016**, *28* (34), 344004. https://doi.org/10.1088/0953-8984/28/34/344004.

260 Sándor, M.; Kiss, R.; Keseru, G. M. Virtual Fragment Docking by Glide: A Validation Study on 190 Protein-Fragment Complexes. *J. Chem. Inf. Model.* **2010**, *50* (6), 1165–1172. https://doi.org/10.1021/ci1000407.

261 Chessari, G.; Buck, I. M.; Day, J. E. H.; Day, P. J.; Iqbal, A.; Johnson, C. N.; Lewis, E. J.; Martins, V.; Miller, D.; Reader, M.; Rees, D. C.; Rich, S. J.; Tamanini, E.; Vitorino, M.; Ward, G. A.; Williams, P. A.; Williams, G.; Wilsher, N. E.; Woolford, A. J. A. Fragment-Based Drug Discovery Targeting Inhibitor of Apoptosis Proteins: Discovery of a Non-Alanine Lead Series with Dual Activity Against CIAP1 and XIAP. *J. Med. Chem.* **2015**, *58* (16), 6574–6588. https://doi.org/10.1021/acs.jmedchem.5b00706.

262 Ndubaku, C.; Varfolomeev, E.; Wang, L.; Zobel, K.; Lau, K.; Elliott, L. O.; Maurer, B.; Fedorova, A. V.; Dynek, J. N.; Koehler, M.; Hymowitz, S. G.; Tsui, V.; Deshayes, K.; Fairbrother, W. J.; Flygare, J. A.; Vucic, D. Antagonism of C-IAP and XIAP Proteins Is Required for Efficient Induction of Cell Death by Small-Molecule IAP Antagonists. *ACS Chem. Biol.* **2009**, *4* (7), 557–566. https://doi.org/10.1021/cb900083m.

263 Li, P.; Song, L. F.; Merz, K. M. Systematic Parameterization of Monovalent Ions Employing the Nonbonded Model. *J. Chem. Theory Comput.* **2015**, *11* (4), 1645–1657. https://doi.org/10.1021/ct500918t.

264 Ballatore, C.; Huryn, D. M.; Smith, A. B. Carboxylic Acid (Bio)Isosteres in Drug Design. *ChemMedChem* **2013**, *8* (3), 385–395. https://doi.org/10.1002/cmdc.201200585.

265 Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095. https://doi.org/10.1021/ct400341p.

266 Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. **1996**.

267 Delano, W. L. PyMOL: An Open-Source Molecular Graphics Tool. *CCP4 Newsl. Protein Crystallogr.* **2002**, *40*, 82–92.

268 R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

269 Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research*. Oxford University Press January 1, 2000, pp 235–242. https://doi.org/10.1093/nar/28.1.235.

270 Hoog, S. S.; Zhao, B.; Winbome, E.; Fisher, S.; Green, D. W.; DesJarlais, R. L.; Newlander, K. A.; Callahan, J. F.; Moore, M. L.; Huffman, W. F.; Abdel-Meguid, S. S. A Check on Rational Drug Design: Crystal Structure of a Complex of Human Immunodeficiency Virus Type 1 Protease with a Novel γ-Turn Mimetic Inhibitor. *J. Med. Chem.* **1995**, *38* (17), 3246–3252. https://doi.org/10.1021/jm00017a008.

271 Robbins, A. H.; Coman, R. M.; Bracho-Sanchez, E.; Fernandez, M. A.; Gilliland, C. T.; Li, M.; Agbandje-McKenna, M.; Wlodawer, A.; Dunn, B. M.; McKenna, R. Structure of the Unbound Form of HIV-1 Subtype A Protease: Comparison with Unbound Forms of Proteases from Other HIV Subtypes. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (3), 233–242. https://doi.org/10.1107/S0907444909054298.

272 Smith, B. J.; Colman, P. M.; Von Itzstein, M.; Danylec, B.; Varghese, J. N. Analysis of Inhibitor Binding in Influenza Virus Neuraminidase. *Protein Sci.* **2001**, *10* (4), 689–696. https://doi.org/10.1110/ps.41801.

273 Streltsov, V. A.; Schmidta, P. M.; Breschkina, J. L. M. K. Structure of an Influenza a Virus N9 Neuraminidase with a Tetrabrachion-Domain Stalk. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2019**, *75* (2), 89–97. https://doi.org/10.1107/S2053230X18017892.

274 Vitale, R. M.; Alterio, V.; Innocenti, A.; Winum, J. Y.; Monti, S. M.; De Simone, G.; Supuran, C. T. Carbonic Anhydrase Inhibitors. Comparison of Aliphatic Sulfamate/Bis- Sulfamate Adducts with Isozymes II and IX as a Platform for Designing Tight-Binding, More Isoform-Selective Inhibitors. *J.*

Med. Chem. **2009**, *52* (19), 5990–5998. https://doi.org/10.1021/jm900641r.

275 Eriksson, A. E.; Jones, T. A.; Liljas, A. Refined Structure of Human Carbonic Anhydrase II at 2.0 Å Resolution. *Proteins Struct. Funct. Bioinforma.* **1988**, *4* (4), 274–282. https://doi.org/10.1002/prot.340040406.

276 Anselm, L.; Banner, D. W.; Benz, J.; Groebke Zbinden, K.; Himber, J.; Hilpert, H.; Huber, W.; Kuhn, B.; Mary, J. L.; Otteneder, M. B.; Panday, N.; Ricklin, F.; Stahl, M.; Thomi, S.; Haap, W. Discovery of a Factor Xa Inhibitor (3R,4R)-1-(2,2-Difluoro-Ethyl)- Pyrrolidine-3,4-Dicarboxylic Acid 3-[(5-Chloro-Pyridin-2-Yl)-Amide] 4-{[2-Fluoro-4-(2-Oxo-2H-Pyridin-1-Yl)-Phenyl]-Amide} as a Clinical Candidate. *Bioorganic Med. Chem. Lett.* **2010**, *20* (17), 5313–5319. https://doi.org/10.1016/j.bmcl.2010.06.126.

277 Padmanabhan, K.; Padmanabhan, K. P.; Tulinsky, A.; Park, C. H.; Bode, W.; Huber, R.; Blankenship, D. T.; Cardin, A. D.; Kisiel, W. Structure of Human Des(1-45) Factor Xa at 2.2 Å Resolution. *J. Mol. Biol.* **1993**, *232* (3), 947–966. https://doi.org/10.1006/jmbi.1993.1441.

278 Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W. Thermodynamic Analysis of Water Molecules at the Surface of Proteins and Applications to Binding Site Prediction and Characterization. *Proteins* **2012**, *80* (3), 871–883. https://doi.org/10.1002/prot.23244.

279 Vukovic, S.; Brennan, P. E.; Huggins, D. J. Exploring the Role of Water in Molecular Recognition: Predicting Protein Ligandability Using a Combinatorial Search of Surface Hydration Sites. *J. Phys. Condens. Matter* **2016**, *28* (34), 344007. https://doi.org/10.1088/0953-8984/28/34/344007.

280 Krasowski, A.; Muthas, D.; Sarkar, A.; Schmitt, S.; Brenk, R. DrugPred: A Structure-Based Approach to Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J. Chem. Inf. Model.* **2011**, *51* (11), 2829–2842. https://doi.org/10.1021/ci200266d.

281 Brodney, M. A.; Barreiro, G.; Ogilvie, K.; Hajos-Korcsok, E.; Murray, J.; Vajdos, F.; Ambroise, C.; Christoffersen, C.; Fisher, K.; Lanyon, L.; Liu, J.; Nolan, C. E.; Withka, J. M.; Borzilleri, K. A.; Efremov, I.; Oborski, C. E.; Varghese, A.; Oneill, B. T. Spirocyclic Sulfamides as β-Secretase 1 (BACE-1) Inhibitors for the Treatment of Alzheimers Disease: Utilization of Structure Based Drug Design, Watermap, and Cns Penetration Studies to Identify Centrally Efficacious Inhibitors. *J. Med. Chem.* **2012**, *55* (21), 9224–9239. https://doi.org/10.1021/jm3009426.

282 Englert, L.; Biela, A.; Zayed, M.; Heine, A.; Hangauer, D.; Klebe, G. Displacement of Disordered Water Molecules from Hydrophobic Pocket Creates Enthalpic Signature: Binding of Phosphonamidate to the S1'-Pocket of Thermolysin. *Biochim. Biophys. Acta - Gen. Subj.* **2010**, *1800* (11), 1192–1202. https://doi.org/10.1016/j.bbagen.2010.06.009.

283 Biela, A.; Nasief, N. N.; Betz, M.; Heine, A.; Hangauer, D.; Klebe, G. Dissecting the Hydrophobic Effect on the Molecular Level: The Role of Water, Enthalpy, and Entropy in Ligand Binding to Thermolysin. *Angew. Chemie - Int. Ed.* **2013**, *52* (6), 1822–1828. https://doi.org/10.1002/anie.201208561.

284 Krimmer, S. G.; Cramer, J.; Schiebel, J.; Heine, A.; Klebe, G. How Nothing Boosts Affinity: Hydrophobic Ligand Binding to the Virtually Vacated S1′ Pocket of Thermolysin. *J. Am. Chem. Soc.* **2017**, *139* (30), 10419–10431. https://doi.org/10.1021/jacs.7b05028.

285 Cramer, J.; Krimmer, S. G.; Heine, A.; Klebe, G. Paying the Price of Desolvation in Solvent-Exposed Protein Pockets: Impact of Distal Solubilizing Groups on Affinity and Binding Thermodynamics in a Series of Thermolysin Inhibitors. *J. Med. Chem.* **2017**, *60* (13), 5791–5799. https://doi.org/10.1021/acs.jmedchem.7b00490.

286 Roderick, S. L.; Fournie-Zaluski, M. C.; Roques, B. P.; Matthews, B. W. Thiorphan and Retro-Thiorphan Display Equivalent Interactions When Bound to Crystalline Thermolysin. *Biochemistry* **1989**, *28* (4), 1493–1497. https://doi.org/10.1021/bi00430a011.

287 Gaucher, J. F.; Selkti, M.; Tiraboschi, G.; Prangé, T.; Roques, B. P.; Tomas, A.; Fournié-Zaluski, M. C. Crystal Structures of α-Mercaptoacyldipeptides in the Thermolysin Active Site: Structural Parameters for a Zn Monodentation or Bidentation in Metalloendopeptidases. *Biochemistry* **1999**, *38* (39), 12569–12576. https://doi.org/10.1021/bi991043z.

288 Monzingo, A. F.; Matthews, B. W. Binding of N-Carboxymethyl Dipeptide Inhibitors to Thermolysin Determined by X-Ray Crystallography: A Novel Class of Transition-State Analogues for Zinc Peptidases. *Biochemistry* **1984**, *23* (24), 5724–5729. https://doi.org/10.1021/bi00319a010.

289 Holden, H. M.; Tronrud, D. E.; Weaver, L. H.; Matthews, B. W.; Monzingo, A. F. Slow- and Fast-Binding

Inhibitors of Thermolysin Display Different Modes of Binding: Crystallographic Analysis of Extended Phosphonamidate Transition-State Analogues. *Biochemistry* **1987**, *26* (26), 8542–8553. https://doi.org/10.1021/bi00400a008.

290 Borsi, V.; Calderone, V.; Fragai, M.; Luchinat, C.; Sarti, N. Entropic Contribution to the Linking Coefficient in Fragment Based Drug Design: A Case Study. *J. Med. Chem.* **2010**, *53* (10), 4285–4289. https://doi.org/10.1021/jm901723z.

291 Devel, L.; Garcia, S.; Czarny, B.; Beau, F.; Lajeunesse, E.; Vera, L.; Georgiadis, D.; Stura, E.; Dive, V. Insights from Selective Non-Phosphinic Inhibitors of MMP-12 Tailored to Fit with an S1′ Loop Canonical Conformation. *J. Biol. Chem.* **2010**, *285* (46), 35900–35909. https://doi.org/10.1074/jbc.M110.139634.

292 Bertini, I.; Calderone, V.; Fragai, M.; Giachetti, A.; Loconte, M.; Luchinat, C.; Maletta, M.; Nativi, C.; Yeo, K. J. Exploring the Subtleties of Drug-Receptor Interactions: The Case of Matrix Metalloproteinases. *J. Am. Chem. Soc.* **2007**, *129* (9), 2466–2475. https://doi.org/10.1021/ja065156z.

293 Bertini, I.; Calderone, V.; Cosenza, M.; Fragai, M.; Lee, Y. M.; Luchinat, C.; Mangani, S.; Terni, B.; Turano, P. Conformational Variability of Matrix Metalloproteinases: Beyond a Single 3D Structure. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (15), 5334–5339. https://doi.org/10.1073/pnas.0407106102.

294 Devel, L.; Beau, F.; Amoura, M.; Vera, L.; Cassar-Lajeunesse, E.; Garcia, S.; Czarny, B.; Stura, E. A.; Dive, V. Simple Pseudo-Dipeptides with a P2′ Glutamate: A Novel Inhibitor Family of Matrix Metalloproteases and Other Metzincins. *J. Biol. Chem.* **2012**, *287* (32), 26647–26656. https://doi.org/10.1074/jbc.M112.380782.

295 Rouanet-Mehouas, C.; Czarny, B.; Beau, F.; Cassar-Lajeunesse, E.; Stura, E. A.; Dive, V.; Devel, L. Zinc-Metalloproteinase Inhibitors: Evaluation of the Complex Role Played by the Zinc-Binding Group on Potency and Selectivity. *J. Med. Chem.* **2017**, *60* (1), 403–414. https://doi.org/10.1021/acs.jmedchem.6b01420.

296 Fiaux, H.; Kuntz, D. A.; Hoffman, D.; Janzer, R. C.; Gerber-Lemaire, S.; Rose, D. R.; Juillerat-Jeanneret, L. Functionalized Pyrrolidine Inhibitors of Human Type II α-Mannosidases as Anti-Cancer Agents: Optimizing the Fit to the Active Site. *Bioorganic Med. Chem.* **2008**, *16* (15), 7337–7346. https://doi.org/10.1016/j.bmc.2008.06.021.

297 Kuntz , D. A.; Zhong , W.; Guo, J.; Rose, D. R.; Boons, G.-J. The Molecular Basis of Inhibition of Golgi α-Mannosidase II by Mannostatin A. *ChemBioChem* **2009**, *10* (2), 268–277. https://doi.org/10.1002/cbic.200800538.

298 Terasaka, T.; Kinoshita, T.; Kuno, M.; Nakanishi, I. A Highly Potent Non-Nucleoside Adenosine Deaminase Inhibitor: Efficient Drug Discovery by Intentional Lead Hybridization. *J. Am. Chem. Soc.* **2004**, *126* (1), 34–35. https://doi.org/10.1021/ja038606l.

299 Kim, K. H. Outliers in SAR and QSAR: 3. Importance of Considering the Role of Water Molecules in Protein-Ligand Interactions and Quantitative Structure-Activity Relationship Studies. *J. Comput. Aided. Mol. Des.* **2021**, *35*, 371–396. https://doi.org/10.1007/s10822-021-00377-7.

300 Nazaré, M.; Will, D. W.; Matter, H.; Schreuder, H.; Ritter, K.; Urmann, M.; Essrich, M.; Bauer, A.; Wagner, M.; Czech, J.; Lorenz, M.; Laux, V.; Wehner, V. Probing the Subpockets of Factor Xa Reveals Two Binding Modes for Inhibitors Based on a 2-Carboxyindole Scaffold: A Study Combining Structure-Activity Relationship and X-Ray Crystallography. *J. Med. Chem.* **2005**, *48* (14), 4511–4525. https://doi.org/10.1021/jm0490540.

301 Wang, S.; Meades, C.; Wood, G.; Osnowski, A.; Anderson, S.; Yuill, R.; Thomas, M.; Mezna, M.; Jackson, W.; Midgley, C.; Griffiths, G.; Fleming, I.; Green, S.; McNae, I.; Wu, S. Y.; McInnes, C.; Zheleva, D.; Walkinshaw, M. D.; Fischer, P. M. 2-Anilino-4-(Thiazol-5-Yl)Pyrimidine CDK Inhibitors: Synthesis, SAR Analysis, X-Ray Crystallography, and Biological Activity. *J. Med. Chem.* **2004**, *47* (7), 1662–1675. https://doi.org/10.1021/jm0309957.

302 Siu, K. K. W.; Lee, J. E.; Sufrin, J. R.; Moffatt, B. A.; McMillan, M.; Cornell, K. A.; Isom, C.; Howell, P. L. Molecular Determinants of Substrate Specificity in Plant 5′-Methylthioadenosine Nucleosidases. *J. Mol. Biol.* **2008**, *378* (1), 112–128. https://doi.org/10.1016/j.jmb.2008.01.088.

303 Lim, H.; Jin, X.; Kim, J.; Hwang, S.; Shin, K. B.; Choi, J.; Nam, K. Y.; No, K. T. Investigation of Hot Spot Region in XIAP Inhibitor Binding Site by Fragment Molecular Orbital Method. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 1217–1225. https://doi.org/10.1016/j.csbj.2019.08.004.

304 Arkin, M. R.; Tang, Y.; Wells, J. A. Small-Molecule Inhibitors of Protein-Protein Interactions:

Progressing toward the Reality. *Chem. Biol.* **2014**, *21* (9), 1102–1114. https://doi.org/10.1016/j.chembiol.2014.09.001.

305 Škopić, M. K.; Salamon, H.; Bugain, O.; Jung, K.; Gohla, A.; Doetsch, L. J.; Santos, D. Dos; Bhat, A.; Wagner, B.; Brunschweiger, A. Acid- and Au(i)-Mediated Synthesis of Hexathymidine-DNA-Heterocycle Chimeras, an Efficient Entry to DNA-Encoded Libraries Inspired by Drug Structures. *Chem. Sci.* **2017**, *8* (5), 3356–3361. https://doi.org/10.1039/c7sc00455a.

306 Tamanini, E.; Buck, I. M.; Chessari, G.; Chiarparin, E.; Day, J. E. H.; Frederickson, M.; Griffiths-Jones, C. M.; Hearn, K.; Heightman, T. D.; Iqbal, A.; Johnson, C. N.; Lewis, E. J.; Martins, V.; Peakman, T.; Reader, M.; Rich, S. J.; Ward, G. A.; Williams, P. A.; Wilsher, N. E. Discovery of a Potent Nonpeptidomimetic, Small-Molecule Antagonist of Cellular Inhibitor of Apoptosis Protein 1 (CIAP1) and X-Linked Inhibitor of Apoptosis Protein (XIAP). *J. Med. Chem.* **2017**, *60* (11), 4611–4625. https://doi.org/10.1021/acs.jmedchem.6b01877.

307 Oltersdorf, T.; Elmore, S. W.; Shoemaker, A. R.; Armstrong, R. C.; Augeri, D. J.; Belli, B. A.; Bruncko, M.; Deckwerth, T. L.; Dinges, J.; Hajduk, P. J.; Joseph, M. K.; Kitada, S.; Korsmeyer, S. J.; Kunzer, A. R.; Letai, A.; Li, C.; Mitten, M. J.; Nettesheim, D. G.; Ng, S. C.; Nimmer, P. M.; O'Connor, J. M.; Oleksijew, A.; Petros, A. M.; Reed, J. C.; Shen, W.; Tahir, S. K.; Thompson, C. B.; Tomaselli, K. J.; Wang, B.; Wendt, M. D.; Zhang, H.; Fesik, S. W.; Rosenberg, S. H. An Inhibitor of Bcl-2 Family Proteins Induces Regression of Solid Tumours. *Nature* **2005**, *435* (7042), 677–681. https://doi.org/10.1038/nature03579.

308 Dömling, A.; Antuch, W.; Beck, B.; Schauer-Vukašinović, V. Isosteric Exchange of the Acylsulfonamide Moiety in Abbott's Bcl-XL Protein Interaction Antagonist. *Bioorganic Med. Chem. Lett.* **2008**, *18* (14), 4115–4117. https://doi.org/10.1016/j.bmcl.2008.05.096.

309 Kunig, V.; Potowski, M.; Gohla, A.; Brunschweiger, A. DNA-Encoded Libraries-an Efficient Small Molecule Discovery Technology for the Biomedical Sciences. *Biological Chemistry*. Walter de Gruyter GmbH June 27, 2018, pp 691–710. https://doi.org/10.1515/hsz-2018-0119.

310 Borchert, J.: Modellierung und Simulation der Bindung verschiedener Liganden an Bcl-xL. Bachelor thesis TU Dortmund University. **2019**

311 Brady, R. M.; Vom, A.; Roy, M. J.; Toovey, N.; Smith, B. J.; Moss, R. M.; Hatzis, E.; Huang, D. C. S.; Parisot, J. P.; Yang, H.; Street, I. P.; Colman, P. M.; Czabotar, P. E.; Baell, J. B.; Lessene, G. De-Novo Designed Library of Benzoylureas as Inhibitors of BCL-XL: Synthesis, Structural and Biochemical Characterization. *J. Med. Chem.* **2014**, *57* (4), 1323–1343. https://doi.org/10.1021/jm401948b.

312 Lee, E. F.; Czabotar, P. E.; Smith, B. J.; Deshayes, K.; Zobel, K.; Colman, P. M.; Fairlie, W. D. Crystal Structure of ABT-737 Complexed with Bcl-XL: Implications for Selectivity of Antagonists of the Bcl-2 Family [1]. *Cell Death and Differentiation*. Nature Publishing Group September 15, 2007, pp 1711–1713. https://doi.org/10.1038/sj.cdd.4402178.

313 Lee, E. F.; Dewson, G.; Evangelista, M.; Pettikiriarachchi, A.; Gold, G. J.; Zhu, H.; Colman, P. M.; Fairlie, W. D. The Functional Differences between Pro-Survival and pro-Apoptotic b Cell Lymphoma 2 (Bcl-2) Proteins Depend on Structural Differences in Their Bcl-2 Homology 3 (BH3) Domains. *J. Biol. Chem.* **2014**, *289* (52), 36001–36017. https://doi.org/10.1074/jbc.M114.610758.

314 Kunig, V. B. K.; Potowski, M.; Akbarzadeh, M.; Klika Škopić, M.; Santos Smith, D.; Arendt, L.; Dormuth, I.; Adihou, H.; Andlovic, B.; Karatas, H.; Shaabani, S.; Zarganes-Tzitzikas, T.; Neochoritis, C. G.; Zhang, R.; Groves, M.; Guéret, S. M.; Ottmann, C.; Rahnenführer, J.; Fried, R.; Dömling, A.; Brunschweiger, A. TEAD–YAP Interaction Inhibitors and MDM2 Binders from DNA-Encoded Indole-Focused Ugi Peptidomimetics. *Angew. Chemie* **2020**, *132* (46), 20518–20522. https://doi.org/10.1002/ange.202006280.

315 Furet, P.; Salem, B.; Mesrouze, Y.; Schmelzle, T.; Lewis, I.; Kallen, J.; Chène, P. Structure-Based Design of Potent Linear Peptide Inhibitors of the YAP-TEAD Protein-Protein Interaction Derived from the YAP Omega-Loop Sequence. *Bioorganic Med. Chem. Lett.* **2019**, *29* (16), 2316–2319. https://doi.org/10.1016/j.bmcl.2019.06.022.

316 Kaan, H. Y. K.; Sim, A. Y. L.; Tan, S. K. J.; Verma, C.; Song, H. Targeting YAP/TAZ-TEAD Protein-Protein Interactions Using Fragment-Based and Computational Modeling Approaches. *PLoS One* **2017**, *12* (6), e0178381. https://doi.org/10.1371/journal.pone.0178381.

7. Appendix

7.1 List of used PDBbind refined set structures

10gs, 184l, 185l, 186l, 187l, 188l, 1a28, 1a4k, 1a4r, 1a4w, 1a69, 1a99, 1a9m, 1a9q, 1aaq, 1add, 1adl, 1ado, 1ai4, 1ai5, 1ai7, 1aid, 1ajn, 1ajp, 1ajq, 1ajv, 1ajx, 1alw, 1amk, 1amw, 1atl, 1avn, 1ax0, 1azm, 1b55, 1b57, 1b6k, 1b6l, 1b8n, 1b8o, 1b8y, 1bcd, 1bcu, 1bdq, 1bgq, 1bhx, 1bju, 1bjv, 1bm7, 1bma, 1bn1, 1bn3, 1bn4, 1bnn, 1bnq, 1bnt, 1bnu, 1bnv, 1bnw, 1bp0, 1bq4, 1br6, 1bty, 1bv7, 1bv9, 1bwa, 1bxo, 1bxq, 1bzc, 1bzj, 1bzy, 1c1r, 1c1u, 1c1v, 1c3x, 1c4u, 1c5c, 1c5n, 1c5o, 1c5p, 1c5q, 1c5s, 1c5t, 1c5x, 1c5y, 1c70, 1c83, 1c86, 1c87, 1c88, 1cbx, 1ceb, 1cet, 1cgl, 1ciz, 1cnw, 1cnx, 1cny, 1cps, 1ctt, 1ctu, 1d09, 1d3d, 1d3p, 1d4h, 1d4i, 1d4j, 1d4k, 1d4l, 1d4p, 1d4y, 1d6v, 1d6w, 1d7i, 1d7j, 1d9i, 1dar, 1det, 1df8, 1dgm, 1dhi, 1dhj, 1dif, 1dl7, 1dmp, 1dqn, 1drj, 1drk, 1drv, 1dud, 1duv, 1dy4, 1dzk, 1e1v, 1e1x, 1e2k, 1e2l, 1e3g, 1e3v, 1e4h, 1e66, 1e6q, 1e6s, 1eb2, 1ebw, 1eby, 1ebz, 1ec0, 1ec1, 1ec2, 1ec3, 1ecq, 1efy, 1egh, 1ejn, 1ela, 1elb, 1elc, 1eld, 1ele, 1enu, 1epo, 1erb, 1ew8, 1ew9, 1ezq, 1f0s, 1f0t, 1f0u, 1f3e, 1f4e, 1f4f, 1f4g, 1f4x, 1f57, 1f5k, 1f5l, 1f73, 1f74, 1f8b, 1f8c, 1f8d, 1f8e, 1fao, 1fjs, 1fkb, 1fkf, 1fkg, 1fkh, 1fki, 1fkw, 1fl3, 1flr, 1fpc, 1fq5, 1ft7, 1ftm, 1fv0, 1fzq, 1g1d, 1g2k, 1g2l, 1g2o, 1g30, 1g32, 1g35, 1g36, 1g45, 1g46, 1g48, 1g4o, 1g52, 1g53, 1g54, 1g74, 1g7f, 1g7g, 1g85, 1gai, 1gar, 1gfy, 1ghw, 1gi1, 1gnm, 1gnn, 1gno, 1gpk, 1gpn, 1grp, 1gvw, 1gwv, 1gx8, 1gyx, 1gyy, 1h1s, 1h22, 1h23, 1h46, 1h4w, 1h6h, 1hbv, 1hdq, 1hee, 1hfs, 1hii, 1hk4, 1hmr, 1hms, 1hmt, 1hn4, 1hos, 1hp5, 1hpo, 1hps, 1hsh, 1hsl, 1hvh, 1hvi, 1hvj, 1hvk, 1hvl, 1hvs, 1hxb, 1hxw, 1hyo, 1i1e, 1i2s, 1i37, 1i5r, 1i7z, 1i9n, 1i9p, 1ie9, 1if7, 1if8, 1igb, 1igj, 1iih, 1iiq, 1ik4, 1ikt, 1ivp, 1iy7, 1izh, 1izi, 1j01, 1j14, 1j16, 1j17, 1j36, 1j37, 1j4r, 1jak, 1jao, 1jaq, 1jgl, 1jmg, 1jqy, 1jsv, 1jvu, 1jys, 1jzs, 1k1i, 1k1j, 1k1l, 1k1n, 1k1o, 1k1y, 1k21, 1k22, 1k27, 1k4g, 1k4h, 1k6c, 1k6p, 1k6t, 1k6v, 1k9s, 1kav, 1kc7, 1kdk, 1koj, 1kpm, 1ksn, 1kv1, 1kv5, 1kyv, 1kzk, 1kzn, 1l83, 1l8g, 1lag, 1lah, 1lbk, 1lcp, 1lee, 1lf2, 1lgw, 1lhu, 1li2, 1li3, 1li6, 1lke, 1lnm, 1lpg, 1lpk, 1lpz, 1lrh, 1lst, 1lyx, 1lzq, 1m0b, 1m0o, 1m0q, 1m1b, 1m2p, 1m2q, 1m2r, 1m2x, 1m48, 1m5w, 1m7y, 1mai, 1mes, 1met, 1mfi, 1mh5, 1mjj, 1mmq, 1mmr, 1moq, 1mq5, 1mq6, 1mrn, 1mrw, 1mrx, 1msm, 1msn, 1mtr, 1mu6, 1mu8, 1mue, 1my4, 1n0s, 1n1m, 1n3i, 1n46, 1n4h, 1n5r, 1nc1, 1nc3, 1ndv, 1ndw, 1ndy, 1ndz, 1nf8, 1nfu, 1nfw, 1nfx, 1nfy, 1njc, 1nje, 1njs, 1nki, 1nli, 1nm6, 1no6, 1np0, 1nq7, 1nt1, 1nvq, 1nvr, 1nvs, 1nw4, 1nw5, 1nw7, 1o0h, 1o0m, 1o0n, 1o1s, 1o2h, 1o2r, 1o3i, 1o5a, 1o5c, 1o5g, 1o5r, 1o7o, 1o86, 1oar, 1oba, 1od8, 1odi, 1odj, 1oe8, 1ogd, 1ogx, 1ogz, 1ohr, 1oif, 1okl, 1om1, 1onz, 1ork, 1os0, 1oss, 1oxr, 1oyq, 1oyt, 1p19, 1p1n, 1p1o, 1p1q, 1p57, 1pa9, 1pb8, 1pb9, 1pbq, 1pdz, 1pfu, 1pgp, 1pkx, 1pme, 1pot, 1ppc, 1pph, 1ppk, 1ppl, 1ppm, 1pr5, 1pro, 1ps3, 1px4, 1pxn, 1pxo, 1pxp, 1pzi, 1pzp, 1q1g, 1q5k, 1q65, 1q72, 1q7a, 1q84, 1q8t, 1q8u, 1q8w, 1q91, 1qan, 1qaw, 1qb1, 1qbq, 1qbr, 1qbs, 1qbv, 1qf0, 1qf1, 1qf2, 1qft, 1qin, 1qji, 1qk3, 1qk4, 1qkt, 1ql7, 1ql9, 1qy1, 1qy2, 1qyg, 1r0p, 1r1h, 1r1j, 1r5y, 1r9l, 1rbp, 1rd4, 1rjk, 1rmz, 1rnm, 1rnt, 1ro6, 1rpf, 1rpj, 1rql, 1rr6, 1rtf, 1s19, 1s38, 1s39, 1s5z, 1s89, 1sbg, 1sdt, 1sdu, 1sdv, 1sgu, 1sh9, 1siv, 1sln, 1sqa, 1sqo, 1sqt, 1sr7, 1srg, 1ssq, 1stc, 1sv3, 1sw2, 1swg, 1swr, 1syh, 1syi, 1szd, 1t31, 1t4v, 1t5f, 1t7d, 1tcx, 1td7, 1tjp, 1tkb, 1tlp, 1tmn, 1tng, 1tnh, 1tni, 1tom, 1tx7, 1txr, 1u1w, 1u33, 1u71, 1uho, 1ui0, 1uj5, 1uml, 1uou, 1upf, 1usi, 1usk, 1usn, 1utj, 1utl, 1utm, 1utn, 1uto, 1uv6, 1uvt, 1uw6, 1uwf, 1uwt, 1uwu, 1uz1, 1uz4, 1v11, 1v16, 1v1j, 1v1m, 1v2j, 1v2k, 1v2l, 1v2n, 1v2o, 1v2r, 1v2s, 1v2t, 1v2u, 1v2w, 1v48, 1v7a, 1vfn, 1vso, 1vyf, 1vyg, 1vzq, 1w0z, 1w11, 1w13, 1w3j, 1w4o, 1w4p, 1w4q, 1w5v, 1w5w, 1w5x, 1w5y, 1w7g, 1w96, 1wcq, 1wht, 1wm1, 1wn6, 1ws1, 1ws4, 1wvj, 1x1z, 1x38, 1x39, 1x8d, 1x8j, 1xap, 1xbo, 1xd0, 1xh5, 1xhy, 1xjd, 1xk5, 1xk9, 1xka, 1xkk, 1xow, 1xpz, 1xq0, 1xt8, 1xug, 1xws, 1y20, 1y3v, 1y3x, 1y6q, 1y6r, 1yc1, 1yc4, 1yda, 1ydb, 1ydd, 1ydk, 1ydr, 1yds, 1ydt, 1yet, 1yfz, 1yp9, 1ype, 1ypg, 1ypj, 1yq7, 1yqj, 1yvm, 1z1h, 1z6e, 1z71, 1z9g, 1z9y, 1zc9, 1zdp, 1zfq, 1zge, 1zhy, 1zoe, 1zog, 1zoh, 1zp8, 1zpa, 1zs0, 1zsf, 1zvx, 2a14, 2a4m, 2a5b, 2a5c, 2aac, 2afw, 2afx, 2aj8, 2al5, 2am4, 2amt, 2ans, 2aoc, 2aod, 2aoe, 2aog, 2aqu, 2arm, 2avm, 2avo, 2avq, 2avs, 2ayr, 2b1g, 2b4l, 2b7d, 2b9a, 2baj, 2bak, 2bal, 2bet, 2bfq, 2bmk, 2bo4, 2boh, 2boj, 2bpv, 2bpy, 2bq7, 2bqv, 2br1, 2brb, 2brm, 2bt9, 2bvd, 2bvr, 2bvs, 2byr, 2bys, 2bza, 2c3i, 2c3l, 2c80, 2c94, 2ca8, 2cbj, 2cbu, 2cbv, 2cc7, 2ccb, 2ccc, 2cej, 2cen, 2ceq, 2cer, 2ces, 2cet, 2cex, 2cf8, 2cgf, 2cgr, 2cht, 2cle, 2cli, 2clk, 2cn0, 2csn, 2ctc, 2d0k, 2d1n, 2d1o, 2d3u, 2d3z, 2doo, 2drc, 2dri, 2dw7, 2e1w, 2e27, 2e2r, 2e7f, 2e9u, 2epn, 2erz, 2ewa, 2ewb, 2exm, 2ez7, 2f1g, 2f34, 2f35, 2f7i, 2f7o, 2f7p, 2f80, 2f8g, 2fdp, 2fle, 2flr, 2fmb, 2fqo, 2fqt, 2fqw, 2fqx, 2fqy, 2fu8, 2fvd, 2fw6, 2fxs, 2fxu, 2fxv, 2g5u, 2g94, 2gh9, 2gj5, 2gkl, 2gl0, 2glp, 2gss, 2gst, 2gv7, 2gvv, 2gyi, 2gzl, 2h15, 2h21, 2h3e, 2h4n, 2h6b, 2ha2, 2ha3, 2ha6, 2hah, 2hb3, 2hhn, 2hjb, 2hl4, 2hnc, 2hnx, 2hoc, 2hu6, 2hzl, 2hzy, 2i0a, 2i2c, 2i4j, 2i4u, 2i4v, 2i4w, 2i4x, 2i4z, 2i6b, 2i80, 2ihj, 2ihq, 2iuz, 2iwx, 2izl, 2j27, 2j2u, 2j34, 2j47, 2j4g, 2j4i, 2j62, 2j75, 2j77, 2j78, 2j79, 2j7b, 2j7d, 2j7e, 2j7f, 2j7g, 2j7h, 2j94, 2j95, 2jdm, 2jdp, 2jds, 2jdu, 2jf4, 2jfz, 2jg0, 2jgs, 2jh0, 2jiw, 2jke, 2jkh, 2jkp, 2jxr, 2mas, 2nmx, 2nmz, 2nn1, 2nn7, 2nnd, 2nsj, 2nsl, 2nt7, 2o0u, 2o4j, 2o4k, 2o4l, 2o4n, 2o4r, 2o4z, 2oag, 2oax, 2oc2, 2ogy, 2oi2, 2oiq, 2ojg, 2ojj, 2olb, 2ole, 2on6, 2ovv, 2ovy, 2oxd, 2oxn, 2oxx, 2oxy, 2oym, 2p15, 2p16, 2p2a, 2p3a, 2p3b, 2p3c, 2p3i, 2p4j, 2p4s, 2p4y, 2p53, 2p7a, 2p7g, 2p7z, 2p95, 2pbw, 2pcp, 2pgz, 2pk6, 2pog, 2pou, 2pov, 2pow, 2pql, 2pqz, 2psu, 2psv, 2ptz, 2pu2, 2pvh, 2pvj, 2pvk, 2pvm, 2pvu, 2pwc, 2pwd, 2pwg, 2pwr, 2pym, 2pyn, 2q1q, 2q38, 2q54, 2q55, 2q5k, 2q63, 2q64, 2q7q, 2q88, 2q89, 2q8h, 2q8z, 2qbq, 2qbu, 2qdt, 2qe4, 2qg0, 2qg2, 2qhy, 2qhz, 2qi0, 2qi1, 2qi3, 2qi4, 2qi5, 2qi6, 2qi7, 2qm9, 2qmg, 2qnn, 2qnp, 2qnq, 2qpq, 2qpu, 2qrl, 2qtg, 2qtt, 2qw1, 2qwb, 2qwc, 2qwd, 2qwe, 2qwf, 2qzr, 2r0h, 2r1y, 2r2m, 2r2w, 2r38, 2r3t, 2r3w, 2r43, 2r58, 2r59, 2r5a, 2r5p, 2r9w, 2r9x, 2ra0, 2ra6, 2rcb, 2rd6, 2reg, 2ri9, 2rin, 2rka, 2rkd, 2rke, 2rkf, 2rkg, 2rkm, 2sim, 2std, 2tmn, 2usn, 2uwd, 2uxi, 2uxz, 2uy0, 2uy3, 2uy4, 2uy5, 2uyn, 2uyq, 2uz9, 2v00, 2v2c, 2v2h, 2v2q, 2v2v, 2v3d, 2v3u, 2v57, 2v58, 2v59, 2v77, 2v95, 2vb8, 2vba, 2vc9, 2ves, 2vh0, 2vh6, 2vj8, 2vjx, 2vk2, 2vk6, 2vkm, 2vl4, 2vmc, 2vmd, 2vmf, 2vnp, 2vnt, 2vo5, 2vot, 2vpn, 2vpo, 2vqt, 2vrj, 2vuk, 2vvc, 2vvn, 2vvs, 2vvu, 2vvv, 2vw1, 2vw5, 2vwc, 2vwl, 2vwm, 2vwn, 2vwo, 2vyt, 2vzr, 2w26, 2w47, 2w4x, 2w66, 2w67, 2w8j, 2w8w, 2w9h, 2wb5, 2wbg, 2wc3, 2wc4, 2wca, 2web, 2wec, 2wed, 2weg, 2weh, 2wej, 2weo, 2weq, 2wer, 2wf5, 2wgj, 2whp, 2wk6, 2wky, 2wkz, 2wl0, 2wn9, 2wnc, 2wnj, 2wor, 2wq5, 2wr8, 2wuf, 2wvt, 2wvz, 2wzf, 2wzs, 2x00, 2x09, 2x0y, 2x2r, 2x4z, 2x6x, 2x7t, 2x7u, 2x8z, 2x91, 2x95, 2x96, 2x97, 2xab, 2xb7, 2xb8, 2xbv, 2xbx, 2xc0, 2xc4, 2xd9, 2xda, 2xde, 2xdk, 2xdl, 2xdx, 2xef, 2xeg, 2xei, 2xej, 2xht, 2xib, 2xii, 2xj1, 2xj7, 2xjg, 2xjj, 2xjx, 2xm1, 2xm2, 2xmy, 2xn3, 2xn5, 2xnb, 2xog, 2xp7, 2xpk, 2xxt, 2xyd, 2xye, 2xyf, 2xys, 2xyt, 2y5f, 2y5g, 2y5h, 2y7x, 2y7z, 2y8c, 2ya6, 2ya7, 2ya8, 2yb0, 2ydw, 2yek, 2yel, 2yfa, 2yfe, 2yfx, 2yge, 2ygf, 2yhw, 2yi0, 2yi7, 2yix, 2yk1, 2yki, 2ymd, 2yme, 2ypi, 2yz3, 2z1w, 2z94, 2za0, 2za5, 2zb1, 2zc9, 2zcs, 2zda, 2zdk, 2zdl, 2zdm, 2zdn, 2zfp, 2zfs, 2zft, 2zgx, 2zjw, 2zmm, 2zq0, 2zq2, 2zwz, 2zx6, 2zx7, 2zx8, 2zxd, 2zxg, 2zym, 3a2o, 3a5y, 3a9i, 3aas, 3aau, 3ag9, 3agl, 3ahn, 3aho, 3aid, 3alt, 3ao2, 3ao5, 3ap4, 3arq, 3arx, 3axz, 3b1m, 3b24, 3b25, 3b26, 3b27, 3b3c, 3b3x, 3b4f, 3b4p, 3b50, 3b65, 3b7i, 3b7j, 3b7r, 3be9, 3bex, 3bft, 3bfu, 3bgb, 3bgc, 3bgq, 3bgs, 3bgz, 3bkk, 3bkl, 3bl0, 3bl1, 3bqc, 3bra, 3brn, 3bu1, 3buf, 3bug, 3buh, 3bva, 3bwj, 3bxe, 3bxg, 3c2o, 3c2r, 3c2u, 3c39, 3c4h, 3cct, 3ccw, 3ccz, 3cd0, 3cd5, 3cda, 3cdb, 3cf8, 3cfn, 3cj2, 3cj4, 3cj5, 3ckb, 3ckz, 3cl0, 3cm2, 3cow, 3coy, 3coz, 3cs7, 3ctt, 3cyx, 3cyz, 3cz1, 3czv, 3d0b, 3d0e, 3d1x, 3d1y, 3d4y, 3d4z, 3d50, 3d51, 3d52, 3d6o, 3d6p, 3d78, 3d7z, 3d83, 3d8w, 3d8z, 3d91, 3d9z, 3da9, 3daz, 3dbu, 3dc3, 3dcc, 3dd0, 3dd8, 3ddf, 3ddg, 3djk, 3djo, 3djp, 3djq, 3djx, 3dk1, 3dnd, 3dne, 3dp4, 3dp9, 3dsz, 3dx1, 3dx2, 3dx3, 3dx4, 3dyo, 3dzt, 3e5a, 3e5u, 3e6y, 3e92, 3e93, 3eax, 3eb1, 3ebh, 3ebi, 3ebl, 3ebo, 3ebp, 3ed0, 3ehx, 3ehy, 3ejp, 3ejq, 3ejr, 3eko, 3ekr, 3ekw, 3ekx, 3el1, 3el4, 3el5, 3el9, 3elc, 3eqr, 3ewc, 3ewj, 3f15, 3f16, 3f17, 3f18, 3f19, 3f1a, 3f33, 3f34, 3f37, 3f3c, 3f5k, 3f5l, 3f68, 3f70, 3f78, 3f7i, 3f8c, 3f8e, 3f8f, 3fat, 3fcq, 3fed, 3fee, 3ffg, 3ffp, 3fh7, 3fhb, 3fjg, 3fl5, 3fqe, 3fql, 3fv1, 3fv2, 3fvk, 3fvl, 3fvn, 3g0e, 3g0i, 3g0w, 3g1v, 3g2y, 3g2z, 3g30, 3g31, 3g32, 3g34, 3g35, 3g5k, 3ga5, 3gba, 3gbe, 3gc4, 3gc5, 3gcp, 3gcs, 3gcu, 3ge7, 3ggu, 3gi4, 3gi5, 3gi6, 3gjw, 3gk1, 3gkz, 3gm0, 3gnw, 3gqz, 3gr2, 3gs6, 3gsm, 3gst, 3gta, 3gtc, 3gv9, 3gvb, 3gx0, 3gy2, 3gy3, 3gy4, 3gy7, 3h1x, 3h30, 3h78, 3h89, 3hb4, 3hcm, 3hek, 3hig, 3hit, 3hk1, 3hkn, 3hkq, 3hkt, 3hku, 3hkw, 3hky, 3hl5, 3hl7, 3hl8, 3hll, 3hmo, 3hmp, 3hp9, 3hs4, 3hu3, 3huc, 3hv8, 3i25, 3i4b, 3i5z, 3i60, 3i6o, 3i7e, 3i9g, 3ibi, 3ibl, 3ibn, 3ibu, 3ies, 3igp, 3imc, 3ime, 3iob, 3ioc, 3ioe, 3iof, 3ip9, 3iph, 3ipu, 3isj, 3iss, 3iub, 3iue, 3ivc, 3ivg, 3ivx, 3iw5, 3iw6, 3iww, 3jdw, 3jrs, 3jrx, 3juk, 3juo, 3jup, 3jvr, 3jy0, 3jya, 3jyr, 3jzj, 3k00, 3k02, 3k2f, 3k37, 3k4d, 3k5v, 3k5x, 3k8q, 3k97, 3k99, 3kdc, 3kdd, 3kdm, 3kek, 3kgp, 3kgq, 3kgt, 3kgu, 3kiv, 3kjd, 3kku, 3kmc, 3kmx, 3kmy, 3kqr, 3kr4, 3kr8, 3kv2, 3kyq, 3l3m, 3l3n, 3l4u, 3l4w, 3l59, 3ldp, 3ldq, 3le9, 3lea, 3lgs, 3lir, 3liw, 3ljg, 3ljo, 3ljz, 3lk8, 3lka, 3lp4, 3lp7, 3lpi, 3lpk, 3lpl, 3lvw, 3lxe, 3lxk, 3lzs, 3m1k, 3m35, 3m36, 3m37, 3m3z, 3m5e, 3m67, 3m6r, 3m8u, 3m96, 3mam, 3mdz, 3mf5, 3mfv, 3mfw, 3mhc, 3mho, 3mhw, 3mi3, 3miy, 3mjl, 3ml2, 3ml5, 3mmf, 3mna, 3mof, 3ms9, 3mss, 3muz, 3mxd, 3mxe, 3myg, 3myq, 3mzc, 3n0n, 3n2u, 3n2v, 3n35, 3n3g, 3n3j, 3n4b, 3n76, 3n7a, 3n7o, 3n86, 3n8k, 3nb5, 3nee, 3neo, 3nes, 3nex, 3ng4, 3nhi, 3nht, 3nkk, 3nox, 3npc, 3nq3, 3nq9, 3nw3, 3nx7, 3nxq, 3nyx, 3nzk, 3o4k, 3o56, 3o5x, 3o75, 3o7u, 3o84, 3o8p, 3o99, 3o9a, 3o9d, 3o9e, 3o9i, 3o9p, 3oaf, 3ocp, 3ocz, 3oil, 3oim, 3ok9, 3oku, 3ovn, 3owj, 3own, 3oy0, 3oyq, 3oyw, 3ozj, 3ozp, 3p17, 3p2e, 3p3g, 3p3s, 3p3t, 3p4v, 3p58, 3p5l, 3p5o, 3p8n, 3p8o, 3p8p, 3p8z, 3pb7, 3pb8, 3pb9, 3pbb, 3pd8, 3pd9, 3pe1, 3pe2, 3pgl, 3pgu, 3pju, 3pn1, 3pn4, 3po1, 3po6, 3ppm, 3ppp, 3ppq, 3ppr, 3prs, 3ps1, 3pwk, 3pww, 3pyy, 3q1x, 3q2j, 3q44, 3q6w, 3q6z, 3q71, 3qaa, 3qbc, 3qdd, 3qfy, 3qfz, 3qgw, 3qgy, 3qlm, 3qox, 3qps, 3qqs, 3qt6, 3qto, 3qtv, 3qwc, 3qx5, 3qxt, 3qxv, 3r16, 3r17, 3r1v, 3r24, 3r4m, 3r4n, 3r4p, 3r6u, 3r7o, 3r88, 3rdo, 3rdq, 3re4, 3rf4, 3rf5, 3rlb, 3rlp, 3rlq, 3rlr, 3rm4, 3rm9, 3roc, 3rr4, 3rsx, 3rt8, 3rtf, 3ru1, 3rux, 3rv8, 3rwp, 3ryj, 3ryv, 3ryx, 3ryy, 3ryz, 3rz0, 3rz1, 3rz5, 3rz7, 3rz8, 3s0b, 3s0d, 3s0e, 3s2v, 3s5y, 3s6t, 3s71, 3s72, 3s73, 3s75, 3s76, 3s77, 3s78, 3sfg, 3sha, 3shc, 3si3, 3si4, 3sio, 3sjf, 3sk2, 3slz, 3spf, 3sr4, 3st5, 3str, 3su0, 3su1, 3su2, 3su3, 3su4, 3su5, 3su6, 3sue, 3suf, 3sug, 3sur, 3sus, 3sut, 3suu, 3suv, 3suw, 3sv2, 3sw8, 3sww, 3sxf, 3t08, 3t0b, 3t1a, 3t1m, 3t2w, 3t3u, 3t5u, 3t60, 3t64, 3t70, 3t82, 3t83, 3t84, 3t85, 3t8v, 3tay, 3tb6, 3td4, 3tf6, 3tfn, 3tfp, 3tfu, 3th9, 3tmk, 3ts4, 3tsk, 3tt4, 3ttm, 3ttp, 3tu7, 3tvc, 3twp, 3tz0, 3tzm, 3u10, 3u5j, 3u5l, 3u6h, 3u81, 3u8j, 3u8k, 3u8l, 3u8n, 3u92, 3u9q, 3ubd, 3ucj, 3udd, 3ueu, 3uev, 3uew, 3uex, 3ug2, 3ui7, 3uil, 3uj9, 3ujc, 3ujd, 3umq, 3uo4, 3uod, 3up2, 3upk, 3usx, 3uu1, 3uug, 3uuo, 3uxd, 3uxk, 3uxl, 3uz5, 3uzj, 3v2n, 3v2p, 3v2q, 3v3q, 3v4t, 3v51, 3v5p, 3v5t, 3v7x, 3vbd, 3vd4, 3vd9, 3veh, 3vf5, 3vh9, 3vha, 3vhc, 3vhd, 3vhk, 3vjc, 3vje, 3vtr, 3vw2, 3w37, 3w5n, 3w9k, 3w9r, 3wgg, 3wha, 3wjw, 3wmc, 3wtn, 3wto, 3wvm, 3wz6, 3wz7, 3wz8, 3wzn, 3x00, 3zbx, 3zc5, 3zcl, 3zdg, 3zdh, 3zi8, 3zj6, 3zk6, 3zll, 3zln, 3zlr, 3zm9, 3zns, 3zpu, 3zqe, 3zso, 3zsq, 3zsx, 3zsy, 3zt2, 3zt3, 3zv7, 3zxz, 3zyu, 3zze, 456c, 4a4q, 4a4v, 4a4w, 4a6b, 4a6c, 4a6s, 4a7i, 4ab9, 4aba, 4abb, 4abd, 4abe, 4abg, 4abh, 4acc, 4aci, 4ad6, 4afg, 4ag8, 4agc, 4agl, 4agm, 4agn, 4ago, 4agp, 4agq, 4ahr, 4ahs, 4ahu, 4ai5, 4aia, 4aj4, 4aje, 4aji, 4ajl, 4alx, 4ap7, 4app, 4aqh, 4arw, 4asd, 4ase, 4asj, 4att, 4auj, 4av4, 4av5, 4avh, 4avi, 4avj, 4avs, 4ax9, 4ayp, 4ayq, 4ayu, 4az5, 4az6, 4azb, 4azc, 4azg, 4azi, 4b0b, 4b1j, 4b2i, 4b2l, 4b32, 4b33, 4b34, 4b35, 4b3c, 4b3d, 4b5d, 4b5s, 4b5t, 4b5w, 4b6o, 4b6p, 4b6r, 4b6s, 4b73, 4b74, 4b76, 4b7j, 4b7r, 4b9k, 4b9z, 4bah, 4bak, 4bam, 4ban, 4bao, 4baq, 4bb9, 4bc5, 4bck, 4bcn, 4bco, 4bcp, 4bcs, 4bf1, 4bf6, 4bi6, 4bi7, 4bj8, 4bks, 4bkt, 4bqg, 4bqh, 4bqs, 4br3, 4bt3, 4bt4, 4bt5, 4btk, 4bup, 4buq, 4c1y, 4c2v, 4c5d, 4c6u, 4c9x, 4ca5, 4ca7, 4ca8, 4cc5, 4cd0, 4ceb, 4cfl, 4cg8, 4cg9, 4cga, 4cgi, 4cig, 4ciw, 4cj4, 4cjp, 4cjq, 4cjr, 4ck3, 4cl6, 4clj, 4cmo, 4cp7, 4cps, 4cpt, 4cpw, 4cpy, 4cpz, 4cr5, 4cr9, 4cra, 4crb, 4crc, 4crf, 4crl, 4csd, 4css, 4cu7, 4cu8, 4cwf, 4cwn, 4cwo, 4cwp, 4cwq, 4cwr, 4cws, 4cwt, 4d1j, 4d3h, 4d8z, 4da5, 4daf, 4db7, 4dbm, 4ddh, 4ddk, 4ddm, 4de0, 4de1, 4de2, 4de5, 4der, 4des, 4det, 4deu, 4dew, 4dff, 4dfg, 4djo, 4djp, 4djq, 4djr, 4dju, 4djv, 4djw, 4djx, 4djy, 4dko, 4dkp, 4dkq, 4dkr, 4dmw, 4do4, 4do5, 4dq2, 4dst, 4dsu, 4duh, 4dv8, 4dzy, 4e0x, 4e1k, 4e3g, 4e4l, 4e4n, 4e5w, 4e6d, 4e6q, 4e70, 4e7r, 4eb8, 4ef6, 4efk, 4efs, 4egk, 4ehz, 4ei4, 4ej8, 4ejl, 4ek9, 4elf, 4elg, 4elh, 4en4, 4eo6, 4eo8, 4eoh, 4eor, 4epy, 4er1, 4erf, 4etz, 4eu0, 4euo, 4ewn, 4exs, 4f09, 4f0c, 4f1l, 4f2w, 4f39, 4f3c, 4f3k, 4f5y, 4f6u, 4f6w, 4f7v, 4f9u, 4f9w, 4f9y, 4fai, 4fcq, 4fev, 4few, 4ffs, 4fht, 4fk6, 4flp, 4fm7, 4fm8, 4fnn, 4fp1, 4fs4, 4fsl, 4fxq, 4fzj, 4g0z, 4g4p, 4g5f, 4g8m, 4g8n, 4g8v, 4g8y, 4g90, 4g95, 4gbd, 4ge1, 4gfm, 4gfo, 4ggz, 4gid, 4gih, 4gii, 4gj2, 4gkh, 4gki, 4gq4, 4gql, 4gqp, 4gqq, 4gqr, 4gr0, 4gr3, 4gr8, 4gu6, 4gu9, 4gue, 4gzp, 4gzt, 4h3f, 4h3j, 4h42, 4h7q, 4h81, 4h85, 4hbm, 4hdb, 4hdf, 4hdp, 4heg, 4hf4, 4hfp, 4hge, 4hj2, 4ht0, 4ht2, 4hu1, 4hw3, 4hwp, 4hws, 4hy1, 4hym, 4hzm, 4i54, 4i5c, 4i71, 4i72, 4i74, 4i7j, 4i7k, 4i7m, 4i7p, 4i8n, 4i8w, 4i8x, 4i8z, 4i9h, 4i9u, 4ibb, 4ibc, 4ibd, 4ibe, 4ibf, 4ibg, 4ibi, 4ibj, 4ibk, 4igt, 4ih3, 4ih5, 4ih6, 4ih7, 4iic, 4iid, 4iie, 4iif, 4ij1, 4ipi, 4ipj, 4ipn, 4ish, 4isi, 4isu, 4itp, 4iue, 4iuo, 4iva, 4ivb, 4ivc, 4ivd, 4iwz, 4j21, 4j22, 4j28, 4j3l, 4j7d, 4j7e, 4j93, 4jal, 4je7, 4jfk, 4jfm, 4jfs, 4jh0, 4jia, 4jkw, 4jn2, 4jsa, 4jss, 4jsz, 4jwk, 4jx9, 4jxs, 4jym, 4jyt, 4jz1, 4jzi, 4k0y, 4k3h, 4k3n, 4k4j, 4k6i, 4k77, 4k7i, 4k7n, 4k7o, 4k9y, 4kao, 4kb9, 4keq, 4kfq, 4km0, 4km2, 4kn0, 4kn1, 4kni, 4knj, 4knm, 4knn, 4ko8, 4kow, 4kp5, 4kp8, 4ks1, 4ks4, 4ksy, 4kwf, 4kwg, 4kwo, 4kxb, 4kyh, 4kyk, 4kz3, 4kz4, 4kz6, 4kz7, 4kzq, 4kzu, 4l19, 4l2l, 4l4v, 4l4z, 4l51, 4l9i, 4lar, 4lbu, 4lch, 4leq, 4lko, 4lkq, 4ll3, 4llj, 4llk, 4llp, 4llx, 4lm2, 4lm3, 4lm4, 4loh, 4loi, 4loo, 4lov, 4loy, 4luz, 4lvt, 4lxz, 4ly1, 4ly9, 4lyw, 4lzr, 4lzs, 4m0e, 4m0f, 4m0r, 4m0y, 4m12, 4m13, 4m14, 4m2r, 4m2u, 4m2v, 4m2w, 4m3p, 4m6u, 4m8e, 4m8h, 4m8x, 4m8y, 4mc6, 4mc9, 4mdn, 4mgd, 4mhy, 4mhz, 4mjp, 4mme, 4mmm, 4mmp, 4mnp, 4mo4, 4mpn, 4mq6, 4mr3, 4mr6, 4mre, 4mrg, 4mrw, 4msc, 4mss, 4muf, 4mul, 4muv, 4n07, 4n5d, 4n6g, 4n6z, 4n7m, 4n8q, 4n9a, 4na9, 4nbk, 4nbl, 4nbn, 4ndu, 4ngm, 4ngn, 4ngp, 4nh7, 4nh8, 4nl1, 4nnr, 4np2, 4np3, 4nra, 4nue, 4nvp, 4nwc, 4nyf, 4o04, 4o05, 4o07, 4o09, 4o0a, 4o0b, 4o0x, 4o0y, 4o2b, 4o2p, 4o3f, 4o61, 4o97, 4o9v, 4oak, 4oc0, 4oc1, 4oc2, 4oc3, 4oc5, 4ocq, 4oct, 4og3, 4ogj, 4oiv, 4oma, 4omj, 4omk, 4ovf, 4ovg, 4ovh, 4owm, 4owv, 4p3h, 4p58, 4p5d, 4p5z, 4p6w, 4p6x, 4pb2, 4pcs, 4pee, 4pf5, 4pft, 4pfu, 4phu, 4pin, 4pmm, 4pnu, 4poh, 4poj, 4pop, 4pow, 4pox, 4pp0, 4pp3, 4pp5, 4pqa, 4psb, 4pum, 4pv5, 4pzv, 4q08, 4q09, 4q0k, 4q19, 4q3t, 4q3u, 4q4o, 4q4p, 4q4q, 4q4r, 4q4s, 4q6d, 4q6e, 4q7p, 4q7s, 4q7v, 4q7w, 4q81, 4q83, 4q87, 4q8x, 4q8y, 4q90, 4q93, 4q99, 4q9o, 4q9y, 4qac, 4qb3, 4qd6, 4qdk, 4qem, 4qer, 4qev, 4qew, 4qf7, 4qf8, 4qf9, 4qgd, 4qge, 4qgi, 4qj0, 4qjw, 4qjx, 4ql1, 4qlk, 4qll, 4qnb, 4qp2, 4qsu, 4qsv, 4qtl, 4qxo, 4qy3, 4qyy, 4r06, 4r0a, 4r4c, 4r4i, 4r4o, 4r4t, 4r59, 4r5a, 4r5b, 4r5t, 4r73, 4r76, 4ra1, 4rak, 4rdn, 4re2, 4re4, 4rfc, 4rfd, 4rfm, 4rfr, 4riv, 4rj8, 4rlt, 4rlu, 4rlw, 4rn4, 4rpn, 4rpo, 4rqk, 4rqv, 4rr6, 4rra, 4rrf, 4rrg, 4rsk, 4rux, 4ruy, 4ruz, 4rvr, 4rwj, 4rww, 4s1g, 4std, 4tjz, 4tkb, 4tkh, 4tkj, 4tmn, 4tqn, 4trc, 4tte, 4tu4, 4tun, 4twp, 4ty6, 4ty7, 4tz2, 4u0f, 4u43, 4u54, 4u5n, 4u5o, 4u5s, 4u6c, 4u70, 4u71, 4u73, 4u8w, 4ua8, 4uac, 4ual, 4ucc, 4ufh, 4ufi, 4ufj, 4ufk, 4ufl, 4ufm, 4uin, 4uj1, 4uj2, 4uja, 4ujb, 4uma, 4und, 4unp, 4up5, 4ury, 4urz, 4uye, 4uyf, 4v01, 4v24, 4w52, 4w97, 4w9c, 4w9d, 4w9f, 4w9h, 4w9i, 4w9j, 4w9k, 4w9l, 4w9o, 4w9p, 4wa9, 4whs, 4wiv, 4wk1, 4wkb, 4wkn, 4wko, 4wkp, 4wn5, 4wov, 4wrb, 4wt2, 4x24, 4x48, 4x50, 4x5p, 4x5r, 4x5y, 4x6m, 4x6n, 4x6o, 4x8o, 4x8u, 4x8v, 4xaq, 4xar, 4xas, 4xip, 4xiq, 4xir, 4xit, 4xk9, 4xmb, 4xo8, 4xoc, 4xoe, 4xt2, 4xtv, 4xty, 4xu0, 4xu1, 4xu2, 4xu3, 4xxh, 4xy8, 4xya, 4y0a, 4y2q, 4y3j, 4y3y, 4y4j, 4y59, 4y5d, 4y79, 4y8x, 4ybk, 4yc0, 4yes, 4ygf, 4yha, 4yhm, 4yho, 4ykk, 4ymb, 4ymg, 4ymh, 4yml, 4ymq, 4ynb, 4yo8, 4yrd, 4ysl, 4ytc, 4yx4, 4yxi, 4yyt, 4yzu, 4z07, 4z0k, 4z0q, 4z1e, 4z1j, 4z1k, 4z83, 4z84, 4z93, 4zae, 4zb6, 4zb8, 4zba, 4zbf, 4zbi, 4zcs, 4zeb, 4zec, 4zek, 4zgk, 4zip, 4zji, 4zl4, 4zls, 4zme, 4zo5, 4zt8, 4zvi, 4zw5, 4zw6, 4zw7, 4zw8, 4zwx, 4zx0, 4zx1, 4zx3, 4zx4, 4zyf, 4zzd, 4zzx, 4zzy, 4zzz, 5a5q, 5a6k, 5a6x, 5a7b, 5a81, 5aa9, 5aan, 5aba, 5acy, 5afv, 5ahw, 5alb, 5am6, 5am7, 5amd, 5amg, 5aml, 5ant, 5anu, 5anv, 5aoi, 5aoj, 5aol, 5aqz, 5aut, 5avf, 5ayt, 5b25, 5b2d, 5b5f, 5b5g, 5boj, 5bry, 5bs4, 5btx, 5bw4, 5bwc, 5byi, 5c1m, 5c1w, 5c28, 5c2a, 5c2h, 5c3p, 5c5t, 5c8n, 5cap, 5caq, 5cas, 5cau, 5cbm, 5cbr, 5cbs, 5cc2, 5cep, 5chk, 5cj6, 5cjf, 5cks, 5cp5, 5cp9, 5cqt, 5cqu, 5cs6, 5cso, 5csp, 5cst, 5cu4, 5cxa, 5cy9, 5czm, 5d0c, 5d0r, 5d1r, 5d21, 5d24, 5d25, 5d26, 5d2r, 5d3c, 5d3h, 5d3j, 5d3l, 5d3n, 5d3p, 5d3t, 5d45, 5d47, 5d48, 5dbm, 5dey, 5dfp, 5dgu, 5dgw, 5dh4, 5dh5, 5dit, 5dkn, 5dlx, 5dnu, 5dpx, 5dq8, 5dqc, 5dqe, 5dqf, 5drr, 5dus, 5duw, 5dwr, 5dx4, 5dxt, 5dyo, 5e13, 5e1s, 5e28, 5e2k, 5e2l, 5e2o, 5e2p, 5e2s, 5e73, 5e74, 5e7n, 5e89, 5edb, 5edc, 5eei, 5eek, 5een, 5ef7, 5efh, 5egm, 5egu, 5eh5, 5eh7, 5eh8, 5ehq, 5ehr, 5ehv, 5ehw, 5ei3, 5eij, 5eis, 5ekm, 5el9, 5elw, 5en3, 5eng, 5ep7, 5epn, 5eq1, 5eqe, 5eqp, 5er2, 5er4, 5etb, 5etj, 5eu1, 5ev8, 5evb, 5evd, 5evk, 5ew0, 5ewa, 5ewk, 5ewy, 5exl, 5exm, 5exn, 5eyr, 5f08, 5f0f, 5f1h, 5f1r, 5f1v, 5f25, 5f2p, 5f60, 5f61, 5f63, 5f8y, 5f9b, 5fck, 5fcz, 5fdc, 5fdi, 5fe6, 5fe7, 5fe9, 5fh7, 5fh8, 5fhm, 5fhn, 5fho, 5fl4, 5fl5, 5fl6, 5flo, 5flq, 5fls, 5flt, 5fnc, 5fnd, 5fnf, 5fng, 5fnr, 5fns, 5fnt, 5fnu, 5fol, 5fs5, 5fsc, 5fsn, 5fso, 5fsy, 5ftg, 5fto, 5fut, 5g17, 5g1a, 5g1z, 5g2b, 5g2g, 5g45, 5g46, 5g4m, 5g4n, 5g4o, 5g57, 5g5f, 5g5v, 5g5z, 5g60, 5g61, 5gj9, 5gja, 5gmh, 5gsa, 5h1t, 5h1u, 5h1v, 5h5f, 5h85, 5h8e, 5h8g, 5h9r, 5ha1, 5hbn, 5hbs, 5hct, 5hcv, 5hcy, 5hi7, 5hrv, 5hrw, 5hrx, 5htl, 5htz, 5hu9, 5hvs, 5hvt, 5hwv, 5hz6, 5hz8, 5hz9, 5i1q, 5i29, 5i2e, 5i2f, 5i3a, 5i3v, 5i3w, 5i3x, 5i7x, 5i7y, 5i80, 5i88, 5i8g, 5i9y, 5ia0, 5ia1, 5ia2, 5ia3, 5ia4, 5ia5, 5ie1, 5igm, 5ih9, 5ihh, 5ii2, 5ijr, 5ikb, 5ime, 5ioz, 5ipc, 5ipj, 5ito, 5itp, 5ivc, 5ive, 5ivv, 5ivy, 5iwg, 5ix0, 5iyy, 5j0d, 5j1r, 5j1x, 5j20, 5j27, 5j2x, 5j3l, 5j64, 5j6a, 5j6l, 5j6m, 5j6n, 5j7q, 5j7w, 5j82, 5j86, 5j8m, 5j8u, 5j8z, 5j9x, 5jfp, 5jfu, 5jg1, 5jgi, 5jgq, 5jhb, 5jhk, 5ji8, 5jop, 5jox, 5jq5, 5js3, 5jsg, 5jsj, 5jss, 5jt9, 5jvi, 5jxn, 5jxq, 5jy3, 5k03, 5k0h, 5k0m, 5k1f, 5k8s, 5k9w, 5ka1, 5ka7, 5ka9, 5kab, 5kad, 5kax, 5kbe, 5kej, 5kh3, 5khm, 5kly, 5km9, 5kma, 5ko1, 5ko5, 5kqx, 5kr2, 5kva, 5kz0, 5l2s, 5l30, 5l3a, 5l4i, 5l4j, 5l4m, 5l7e, 5l7g, 5l7h, 5l8a, 5l8y, 5l9g, 5l9i, 5l9l, 5l9o, 5laq, 5ld8, 5lif, 5ljq, 5ljt, 5llc, 5lle, 5llg, 5llh, 5llo, 5llp, 5lny, 5lom, 5lsg, 5lsh, 5ltn, 5lud, 5lvd, 5lvq, 5lvr, 5lwd, 5lz4, 5lz5, 5lz7, 5m25, 5m28, 5m7s, 5m7u, 5m9w, 5ma7, 5meh, 5mes, 5mg2, 5mge, 5mgf, 5mgj, 5mgk, 5mkr, 5mks, 5mlj, 5mme, 5mmg, 5mn1, 5mnn, 5mnr, 5mo8, 5mod, 5mpk, 5mpn, 5mpz, 5mqe, 5mrb, 5mrm, 5mro, 5mrp, 5msb, 5mwh, 5mwp, 5mwy, 5mxf, 5my8, 5mz8, 5n0d, 5n0e, 5n0f, 5n17, 5n18, 5n1r, 5n1s, 5n1z, 5n24, 5n25, 5n2t, 5n2z, 5n31, 5n34, 5n3v, 5n3y, 5n84, 5n93, 5n9r, 5nap, 5nau, 5nbw, 5ndf, 5ne5, 5nea, 5neb, 5nee, 5ngz, 5nih, 5njz, 5nk2, 5nk3, 5nk4, 5nk6, 5nk7, 5nk8, 5nka, 5nkb, 5nkc, 5nkd, 5nkg, 5nkh, 5nki, 5nlk, 5nn5, 5nn6, 5nvv, 5nvw, 5nvx, 5nw0, 5nw1, 5nw2, 5nwe, 5nxi, 5nxp, 5nxw, 5ny1, 5ny3, 5nya, 5nyh, 5nz4, 5nze, 5nzf, 5nzn, 5o1d, 5o1f, 5o1h, 5o2d, 5o4f, 5o5a, 5o9o, 5o9p, 5o9q, 5o9r, 5o9y, 5oa6, 5odx, 5oei, 5oh2, 5oh3, 5oh4, 5oh7, 5oh9, 5oha, 5oku, 5om2, 5om3, 5om7, 5oot, 5op4, 5op5, 5oq8, 5oqu, 5org, 5orh, 5orj, 5ork, 5orv, 5orw, 5os2, 5os4, 5os5, 5os7, 5os8, 5ose, 5osl, 5oss, 5ost, 5ot8, 5ot9, 5ota, 5otc, 5otr, 5otz, 5ouh, 5ov8, 5owl, 5std, 5sym, 5sz0, 5sz1, 5sz2, 5sz3, 5sz4, 5sz5, 5sz6, 5sz7, 5t19, 5t7s, 5t8o, 5t9u, 5t9w, 5ta2, 5tb6, 5tbe, 5tcj, 5tfx, 5th4, 5ti0, 5tmn, 5tp0, 5tpx, 5tt3, 5tuo, 5tuz, 5twj, 5txy, 5ty9, 5tya, 5u0d, 5u0e, 5u0g, 5u0w, 5u0y, 5u0z, 5u11, 5u12, 5u13, 5u14, 5u28, 5u49, 5u4b, 5u4d, 5u6j, 5u8c, 5uc4, 5ucj, 5ueu, 5uez, 5uf0, 5ufc, 5uff, 5ufp, 5ufr, 5ufs, 5uk8, 5ula, 5uln, 5ulp, 5ult, 5umx, 5umy, 5uoo, 5uov, 5upe, 5upf, 5upj, 5upz, 5uv2, 5uxf, 5v0n, 5v79, 5v7a, 5v82, 5var, 5vb5, 5vb7, 5vcv, 5vcw, 5vd0, 5vd1, 5vd2, 5vd3, 5vgy, 5vih, 5vij, 5vja, 5vkc, 5vl2, 5vm0, 5vo1, 5voj, 5vp9, 5vr8, 5vsf, 5vsj, 5vyy, 5w1e, 5w44, 5wa8, 5wa9, 5wal, 5wbm, 5we9, 5wex, 5wgp, 5wl0, 5wlo, 5wp5, 5wqc, 5wuk, 5wyx, 5wyz, 5x62, 5x74, 5xg5, 5xmx, 5xo7, 5xpi, 5xsr, 5xva, 5xvg, 5y12, 5y13, 5y8y, 5y94, 5ya5, 5yas, 5yh8, 5yhe, 5yhg, 5yj8, 5yjm, 5yl2, 5z5f, 5z7b, 5z7j, 5z99, 5za7, 5za8, 5za9, 5zae, 5zaf, 5zag, 5zaj, 5zc5, 5zkc, 6aqs, 6ayo, 6ayq, 6b1k, 6b4l, 6b4u, 6b59, 6b7a, 6b7b, 6b96, 6b97, 6b98, 6bbx, 6bdy, 6bhv, 6bm5, 6bm6, 6c0s, 6c7q, 6c7w, 6c7x, 6cbf, 6cbg, 6cdj, 6cdl, 6ce6, 6ced, 6ckr, 6cks, 6cpa, 6cpw, 6csp, 6csq, 6csr, 6css, 6cwh, 6cwn, 6d2o, 6d50, 6d55, 6d56, 6d5e, 6d5g, 6d5h, 6d5j, 6d9x, 6dai, 6dak, 6dar, 6dh1, 6dh2, 6dh6, 6dh7, 6dh8, 6dif, 6dil, 6dj1, 6dj2, 6dj5, 6dj7, 6dq4, 6e4a, 6e7j, 6e9a, 6edr, 6eed, 6ei5, 6eif, 6eij, 6eiq, 6eir, 6eis, 6eiz, 6ej2, 6ej3, 6ekq, 6el5, 6eln, 6elo, 6elp, 6en5, 6eog, 6eol, 6ep4, 6epa, 6epy, 6epz, 6eq1, 6eq8, 6eqp, 6equ, 6euw, 6eux, 6evr, 6ex1, 6exi, 6exs, 6ey8, 6ey9, 6eya, 6eyb, 6eyt, 6f1j, 6f1n, 6f20, 6f28, 6f3b, 6f90, 6f92, 6f9g, 6f9u, 6f9v, 6fa4, 6faf, 6fba, 6fe0, 6fe1, 6fgg, 6fhq, 6fmc, 6fmj, 6fnf, 6fng, 6fni, 6fnj, 6fnq, 6fnr, 6fo5, 6fs0, 6fs1, 6ftp, 6ftz, 6fuh, 6fui, 6fuj, 6fyz, 6g34, 6g35, 6g36, 6g37, 6g38, 6g39, 6g3a, 6g3q, 6g3v, 6g98, 6g9i, 6g9u, 6ge7, 6gf9, 6gfs, 6gfz, 6ghh, 6gji, 6gjj, 6gjl, 6gjm, 6gjn, 6gjr, 6gl8, 6gl9, 6gnm, 6gnp, 6gnr, 6gnw, 6gon, 6got, 6guc, 6gue, 6guh, 6guk, 6gvz, 6gw4, 6gwr, 6gzd, 6gzm, 6h29, 6h2t, 6h2z, 6h33, 6h34, 6h36, 6h37, 6h38, 6h5x, 6h8s, 6hai, 6hd6, 6hh3, 6hh5, 6hke, 6hlx, 6hly, 6hpw, 6hqy, 6hrq, 6hsh, 6ht1, 6htg, 6iiu, 6ma2, 6ma3, 6ma4, 6ma5, 6mjf, 6msy, 6std, 6upj, 7std, 7upj, 8a3h, 8cpa, 966c

7.2 List of used PDBbind core set 2013 structures

10gs, 1a30, 1bcu, 1e66, 1f8b, 1f8c, 1f8d, 1gpk, 1h23, 1hfs, 1hnn, 1igj, 1jyq, 1kel, 1lbk, 1lol, 1loq, 1lor, 1mq6, 1n1m, 1n2v, 1nvq, 1o3f, 1o5b, 1os0, 1oyt, 1p1q, 1ps3, 1q8t, 1q8u, 1qi0, 1r5y, 1sln, 1sqa, 1u1b, 1u33, 1uto, 1vso, 1w3k, 1w3l, 1w4o, 1xd0, 1yc1, 1z95, 1zea, 2brb, 2cbj, 2cet, 2d1o, 2d3u, 2fvd, 2g70, 2gss, 2hb1, 2iwx, 2j62, 2j78, 2jdm, 2jdu, 2jdy, 2obf, 2ole, 2p4y, 2pcp, 2pq9, 2qbp, 2qbr, 2qft, 2qmj, 2r23, 2v00, 2v7a, 2vl4, 2vo5, 2vot, 2vvn, 2vw5, 2w66, 2wbg, 2wca, 2weg, 2wtv, 2x00, 2x0y, 2x8z, 2x97, 2xb8, 2xbv, 2xdl, 2xhm, 2xnb, 2xy9, 2y5h, 2yfe, 2yge, 2yki, 2ymd, 2zcq, 2zcr, 2zjw, 2zwz, 2zxd, 3acw, 3ag9, 3ao4, 3b3s, 3b3w, 3b68, 3bfu, 3bkk, 3bpc, 3cyx, 3d4z, 3dd0, 3dxg, 3e93, 3ehy, 3ejr, 3f17, 3f3e, 3fcq, 3fk1, 3fv1, 3g0w, 3g2z, 3gbb, 3gcs, 3ge7, 3gnw, 3gy4, 3huc, 3imc, 3ivg, 3jvs, 3k5v, 3kgp, 3kv2, 3kwa, 3l3n, 3lka, 3mfv, 3mss, 3myg, 3n7a, 3nox, 3nq3, 3nw9, 3oe5, 3ov1, 3owj, 3ozt, 3pe2, 3pww, 3pxf, 3s8o, 3su2, 3su3, 3su5, 3u9q, 3udh, 3ueu, 3uex, 3uri, 3utu, 3vh9, 3zso, 3zsx, 4de1, 4de2, 4des, 4dew, 4djr, 4djv, 4g8m, 4gid, 4gqq, 4tmn

7.3 Druggability data set

Druggable: 1uou, 1e66, 1kzn, 2br1, 1lpz, 1o5r

Undruggable: 1v16, 3jdw, 1kc7, 1mai, 1px4, 1od8, 1d09, 1moq, 1rnt, 1onz, 1jak, 2gyi, 1o86

7.4 TEAD4 Brunschweiger data set

ClC1=CC=C(C(CNCC#C)=CN2)C2=C1

ClC1=CC=C(C=C(CNCC2=CN(CCN3CCCCCC3)N=N2)N4)C4=C1

ClC1=CC=C(C=C(CNCC2=CN(CC3=CN(C=C(C)C=C4)C4=N3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CNCC2=CN(CC3=NC=C(C4=CC=CC=C4)O3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(C)=O)CC2=CN(CCN3CCCCCC3)N=N2)N4)C4=C1

ClC1=CC=C(C=C(CN(C(C)=O)CC2=CN(CC3=CN(C=C(C)C=C4)C4=N3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(C)=O)CC2=CN(CC3=NC=C(C4=CC=CC=C4)O3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(CNC(OC(C)(C)C)=O)=O)CC2=CN(CCN3CCCCCC3)N=N2)N4)C4=C1

ClC1=CC=C(C=C(CN(C(CNC(OC(C)(C)C)=O)=O)CC2=CN(CC3=CN(C=C(C)C=C4)C4=N3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(CNC(OC(C)(C)C)=O)=O)CC2=CN(CC3=NC=C(C4=CC=CC=C4)O3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(CCC(OC(C)(C)C)=O)=O)CC2=CN(CCN3CCCCCC3)N=N2)N4)C4=C1

ClC1=CC=C(C=C(CN(C(CCC(OC(C)(C)C)=O)=O)CC2=CN(CC3=CN(C=C(C)C=C4)C4=N3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(CCC(OC(C)(C)C)=O)=O)CC2=CN(CC3=NC=C(C4=CC=CC=C4)O3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(CN)=O)CC2=CN(CCN3CCCCCC3)N=N2)N4)C4=C1

ClC1=CC=C(C=C(CN(C(CCC(O)=O)=O)CC2=CN(CCN3CCCCCC3)N=N2)N4)C4=C1

ClC1=CC=C(C=C(CN(C(CN)=O)CC2=CN(CC3=NC=C(C4=CC=CC=C4)O3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(CCC(O)=O)=O)CC2=CN(CC3=NC=C(C4=CC=CC=C4)O3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(CN)=O)CC2=CN(CC3=CN(C=C(C)C=C4)C4=N3)N=N2)N5)C5=C1

ClC1=CC=C(C=C(CN(C(CCC(O)=O)=O)CC2=CN(CC3=CN(C=C(C)C=C4)C4=N3)N=N2)N5)C5=C1

CC(C)C1=NC(CN(N=N2)C=C2CNCC3=CC4=CC=C(Cl)C=C4N3)=CS1

CC(C)C1=NC(CN(N=N2)C=C2CN(C(CNC(OC(C)(C)C)=O)=O)CC3=CC4=CC=C(Cl)C=C4N3)=CS1

CC(C)C1=NC(CN(N=N2)C=C2CN(C(C)=O)CC3=CC4=CC=C(Cl)C=C4N3)=CS1

```
CC(C)C1=NC(CN(N=N2)C=C2CN(C(CCC(OC(C)(C)C)=O)=O)CC3=CC4=CC=C(Cl)C=C4N3)=CS1
CC(C)C1=NC(CN(N=N2)C=C2CN(C(CN)=O)CC3=CC4=CC=C(Cl)C=C4N3)=CS1
CC(C)C1=NC(CN(N=N2)C=C2CN(C(C(CCCC)N)=O)CC3=CC4=CC=C(Cl)C=C4N3)=CS1
ClC1=CC=C(C=C(CNCC2=CN(CC(C3=CC=CC=C3NC4=CC(C(F)(F)F)=CC=C4)=O)N=N2)N5)C5=C1
ClC1=CC=C(C=C(CN(C(CNC(OC(C)(C)C)=O)=O)CC2=CN(CC(C3=CC=CC=C3NC4=CC(C(F)(F)F)=CC=C4)=O)N=N2)N5)C5=C1
ClC1=CC=C(C=C(CN(C(C)=O)CC2=CN(CC(C3=CC=CC=C3NC4=CC(C(F)(F)F)=CC=C4)=O)N=N2)N5)C5=C1
ClC1=CC=C(C=C(CN(C(CCC(OC(C)(C)C)=O)=O)CC2=CN(CC(C3=CC=CC=C3NC4=CC(C(F)(F)F)=CC=C4)=O)N=N2)N5)C5=C1
ClC1=CC=C(C=C(CN(C(CN)=O)CC2=CN(CC(C3=CC=CC=C3NC4=CC(C(F)(F)F)=CC=C4)=O)N=N2)N5)C5=C1
ClC1=CC=C(C=C(CN(C(C(CCCC)N)=O)CC2=CN(CC(C3=CC=CC=C3NC4=CC(C(F)(F)F)=CC=C4)=O)N=N2)N5)C5=C1
ClC1=CC=C(C=C(CN(C(C2N(C(C(C(C)C)NC(C)=O)=O)CCC2)=O)CC3=CN(CC4=CN(C=C(C)C=C5)C5=N4)N=N3)N6)C6=C1
ClC1=CC=C(C=C(CN(C(C2N(C(C(C(C)C)NC(C)=O)=O)CCC2)=O)CC3=CN(CC4=NC=C(C5=CC=CC=C5)O4)N=N3)N6)C6=C1
ClC1=CC=C(C=C(CN(C(C2N(C(C(C(C)C)NC(C)=O)=O)CCC2)=O)CC3=CN(CC(C4=CC=CC=C4NC5=CC(C(F)(F)F)=CC=C5)=O)N=N3)N6)C6=C1
O=C(C(C(C)C)NC(C)=O)N1CCCC1C(N(CC2=CN(CC3=CSC(C(C)C)=N3)N=N2)CC4=CC5=CC=C(Cl)C=C5N4)=O
ClC1=CC=C(C=C(CN(C(C2N(C(C(C(C)C)NC(C)=O)=O)CCC2)=O)CC3=CN(CCN4CCCCCC4)N=N3)N5)C5=C1
BrC1=CC=CC(NC2=CC=CC=C2C(N(CC3=CC4=CC=C(Cl)C=C4N3)CC5=CN(CCN6CCCCCC6)N=N5)=O)=C1
BrC1=CC=CC(NC2=CC=CC=C2C(N(CC3=CN(CC4=NC=C(C5=CC=CC=C5)O4)N=N3)CC6=CC7=CC=C(Cl)C=C7N6)=O)=C1
ClC1=CC=C(C=C(CN(C(C2=CC=CC=C2NC3=CC(Br)=CC=C3)=O)CC4=CN(CC5=CN(C=C(C)C=C6)C6=N5)N=N4)N7)C7=C1
CC(C)C1=NC(CN(N=N2)C=C2CN(C(C3=CC=CC=C3NC4=CC(Br)=CC=C4)=O)CC5=CC6=CC=C(Cl)C=C6N5)=CS1
ClC1=CC=C(C=C(CN(C(C(N2C=CC=C2)C3=CC=CC=C3)=O)CC4=CN(CCN5CCCCCC5)N=N4)N6)C6=C1
O=C(C(N1C=CC=C1)C2=CC=CC=C2)N(CC3=CN(CC4=NC=C(C5=CC=CC=C5)O4)N=N3)CC6=CC7=CC=C(Cl)C=C7N6
ClC1=CC=C(C=C(CN(C(C(N2C=CC=C2)C3=CC=CC=C3)=O)CC4=CN(CC5=CN(C=C(C)C=C6)C6=N5)N=N4)N7)C7=C1
ClC1=CC=C(C=C(CN(C(C2=C(NC3=CC=CC(C(F)(F)F)=C3)N=CC=C2)=O)CC4=CN(CCN5CCCCCC5)N=N4)N6)C6=C1
O=C(C1=C(NC2=CC=CC(C(F)(F)F)=C2)N=CC=C1)N(CC3=CN(CC4=NC=C(C5=CC=CC=C5)O4)N=N3)CC6=CC7=CC=C(Cl)C=C7N6
CC(C)C1=NC(CN(N=N2)C=C2CN(C(C(N3C=CC=C3)C4=CC=CC=C4)=O)CC5=CC6=CC=C(Cl)C=C6N5)=CS1
ClC1=CC=C(C=C(CN(C(C2=C(NC3=CC=CC(C(F)(F)F)=C3)N=CC=C2)=O)CC4=CN(CC5=CN(C=C(C)C=C6)C6=N5)N=N4)N7)C7=C1
CC(C)C1=NC(CN(N=N2)C=C2CN(C(C3=C(NC4=CC=CC(C(F)(F)F)=C4)N=CC=C3)=O)CC5=CC6=CC=C(Cl)C=C6N5)=CS1
```

7.5 *Holo* water thermodynamics and ligand affinity - scatterplots

7.5.1 Whole PDBbind refined set

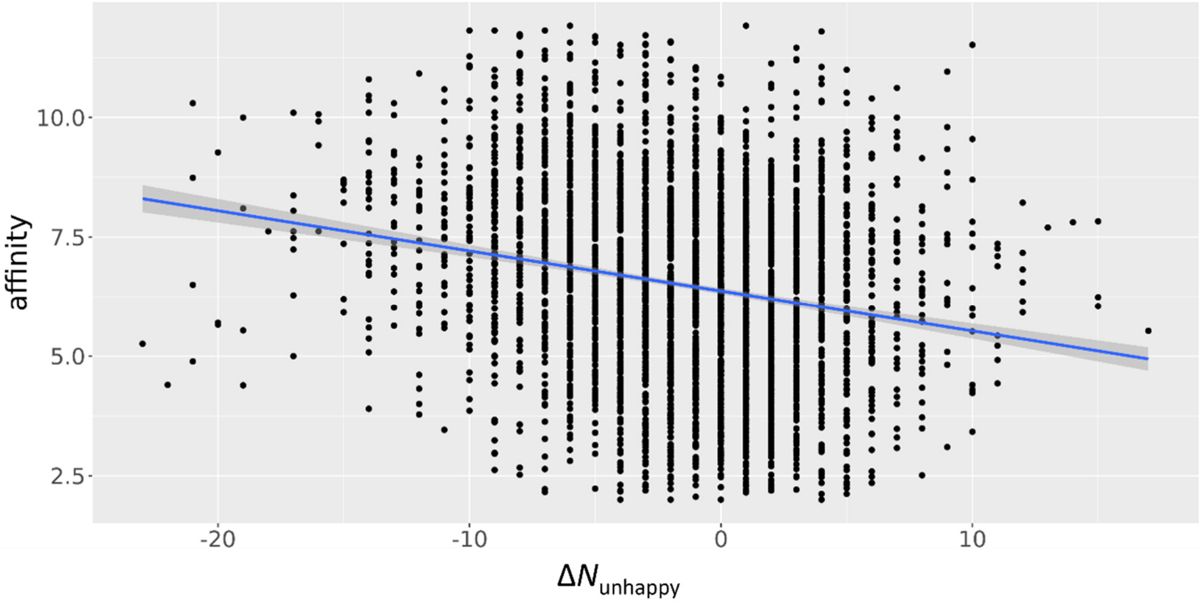

Figure 69: Scatterplot of ligand affinity (pK_i/pK_d) and the parameter $\Delta N_{unhappy}$ as defined in Eq. (71) for complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

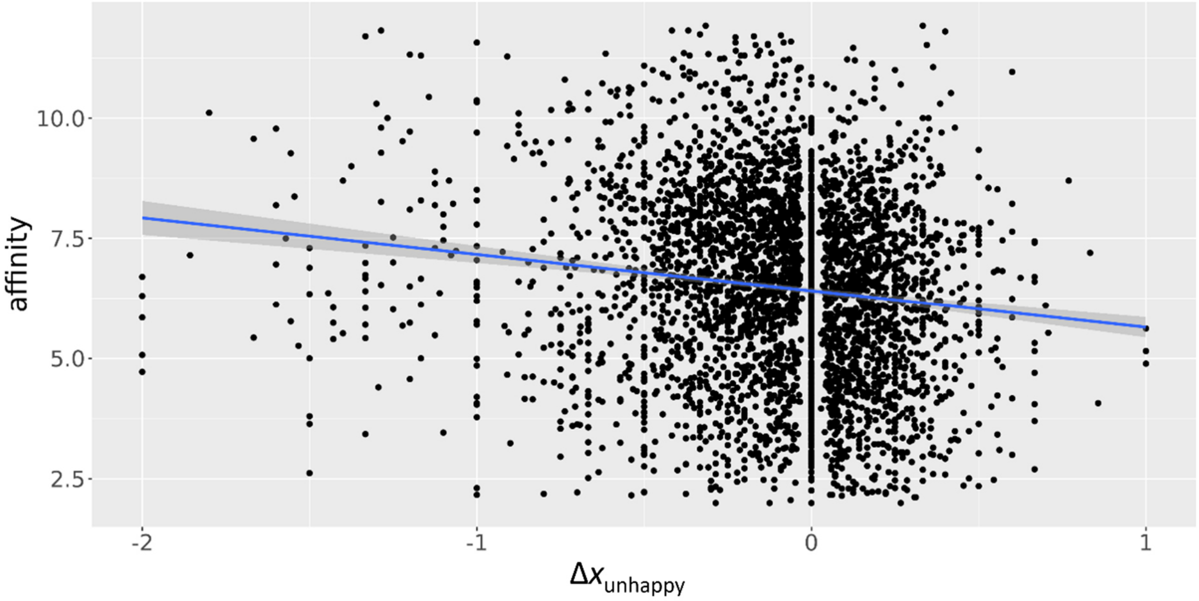

Figure 70: Scatterplot of ligand affinity (pK_i/pK_d) and the parameter $\Delta x_{unhappy}$ as defined in Eq. (72) for complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

7.5.2 Protein subsets

BACE1

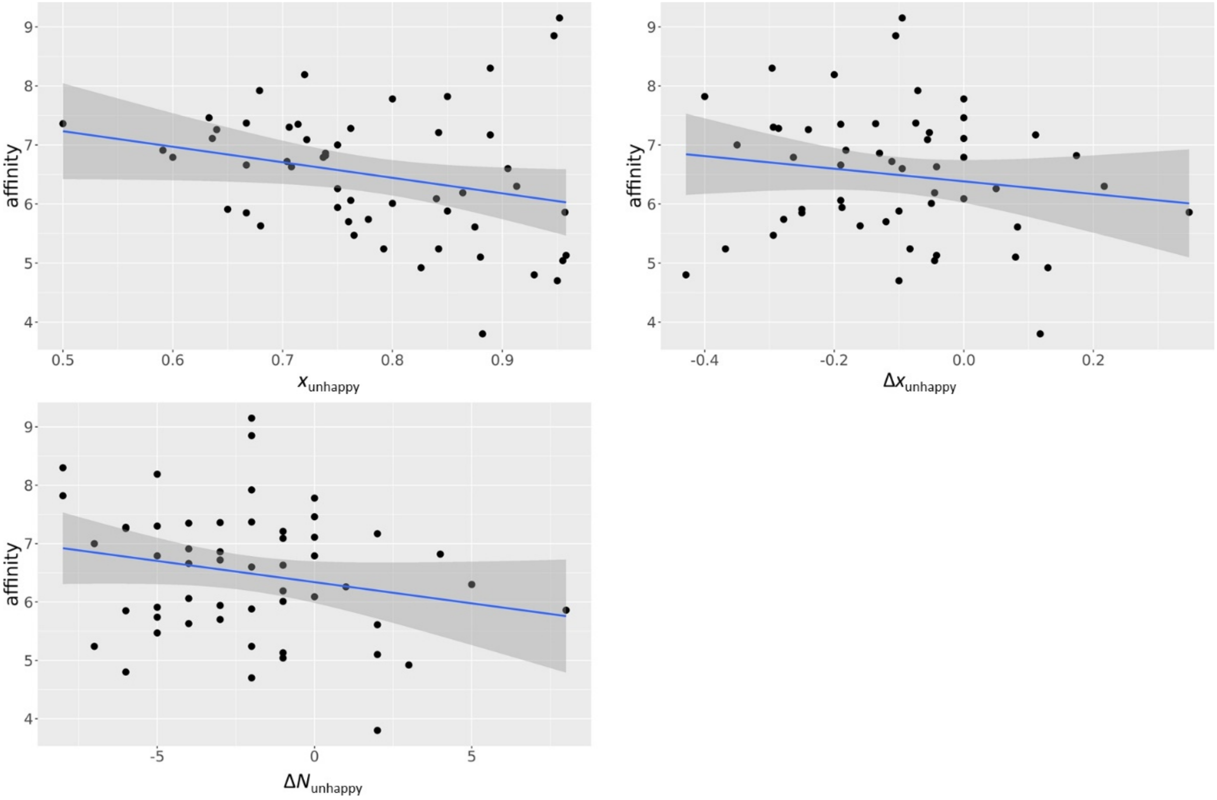

Figure 71: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for BACE1 complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

BRD4

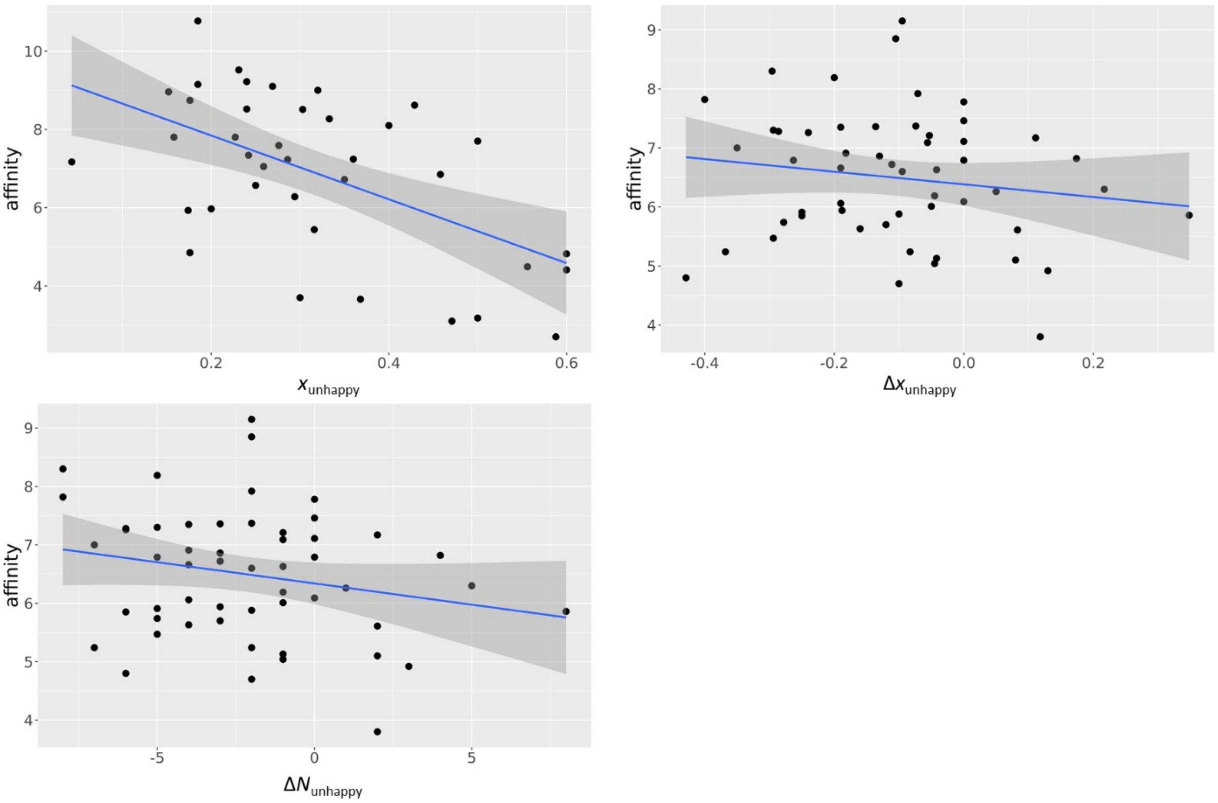

Figure 72: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for BRD4 complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

CA2

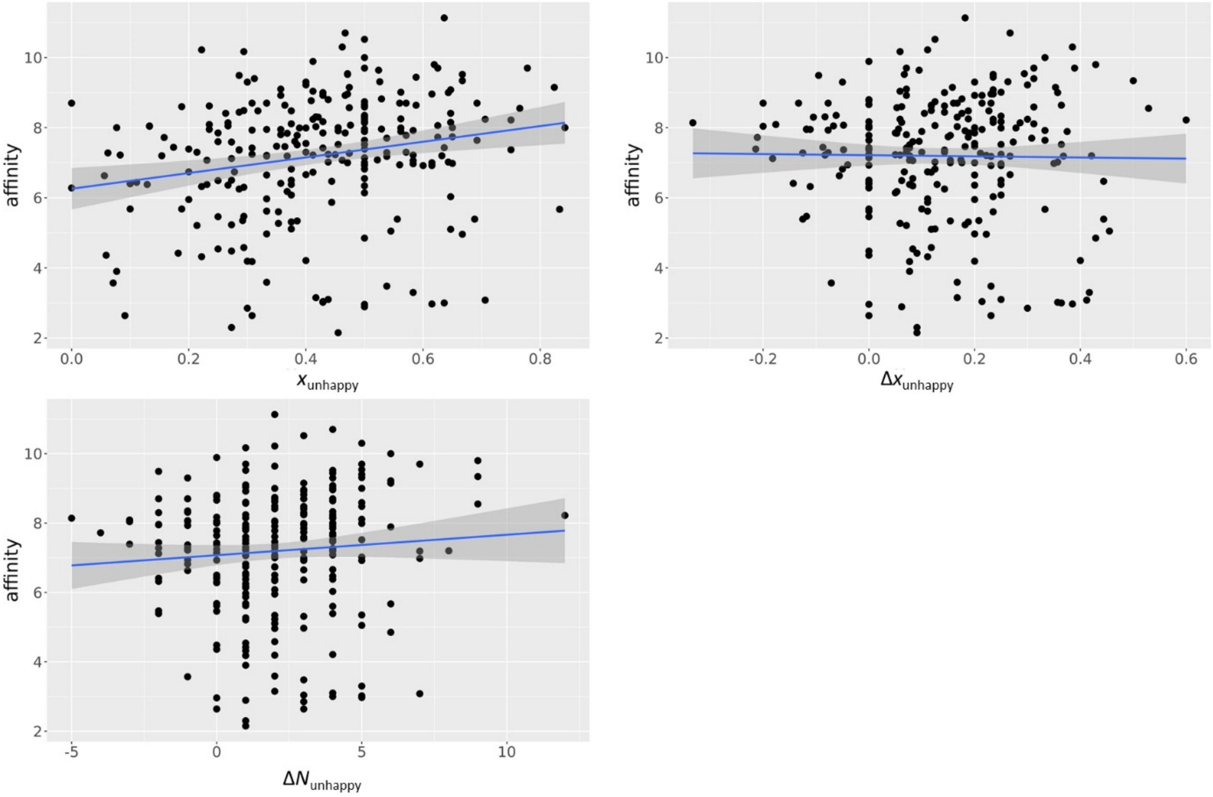

Figure 73: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for CA2 complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

Caseinkinase

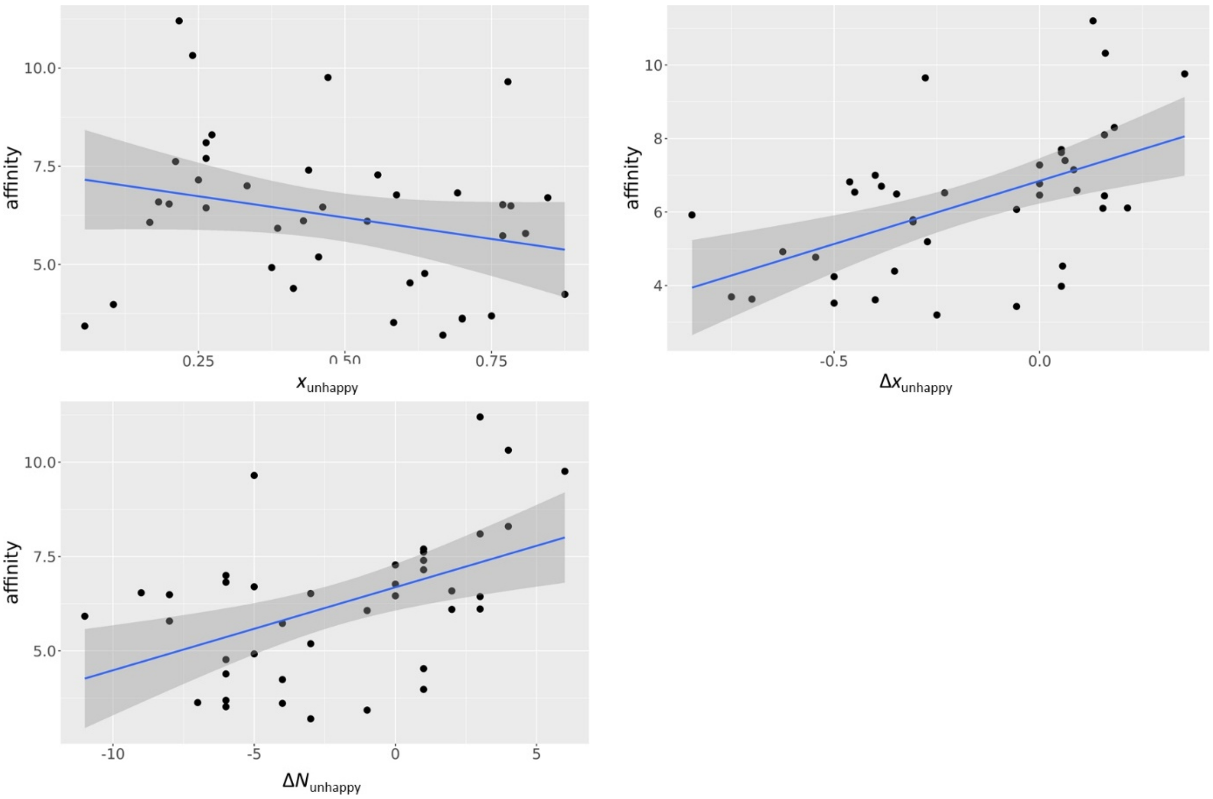

Figure 74: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for Caseinkinase complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

fXa

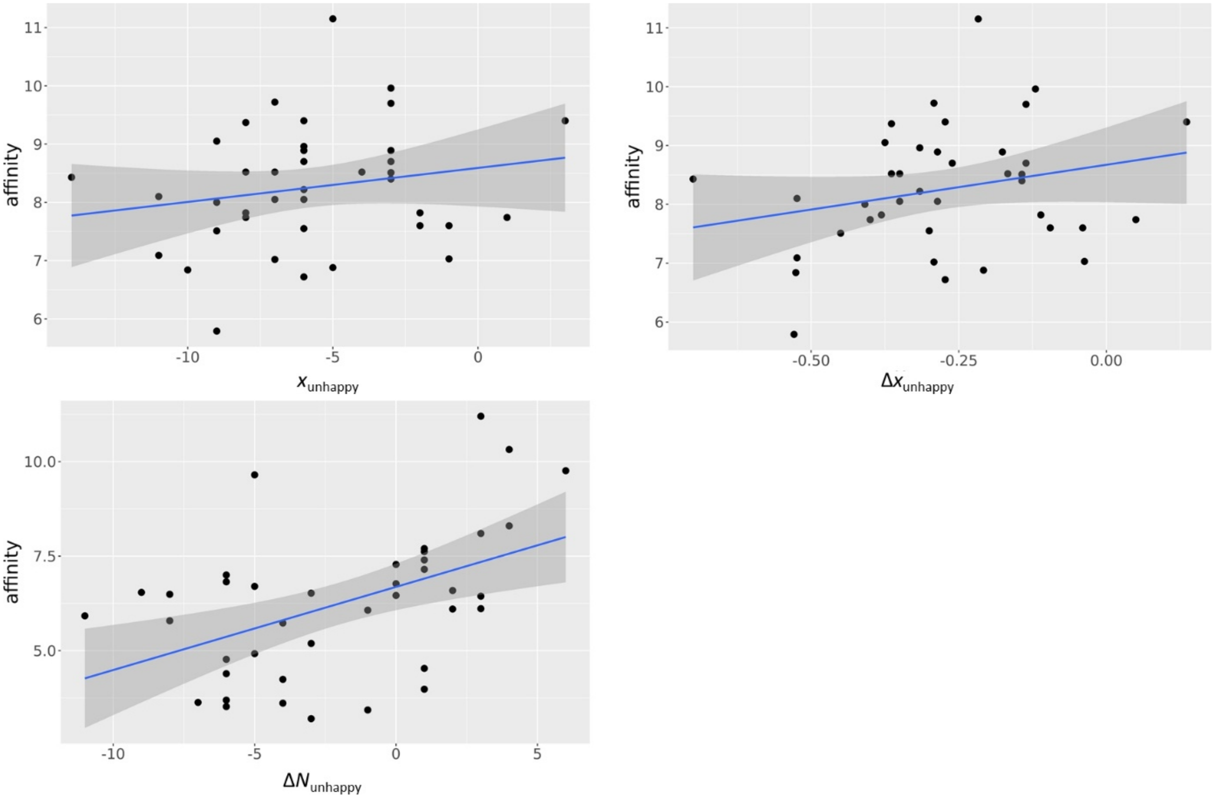

Figure 75: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for fXa complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

HIV1PR

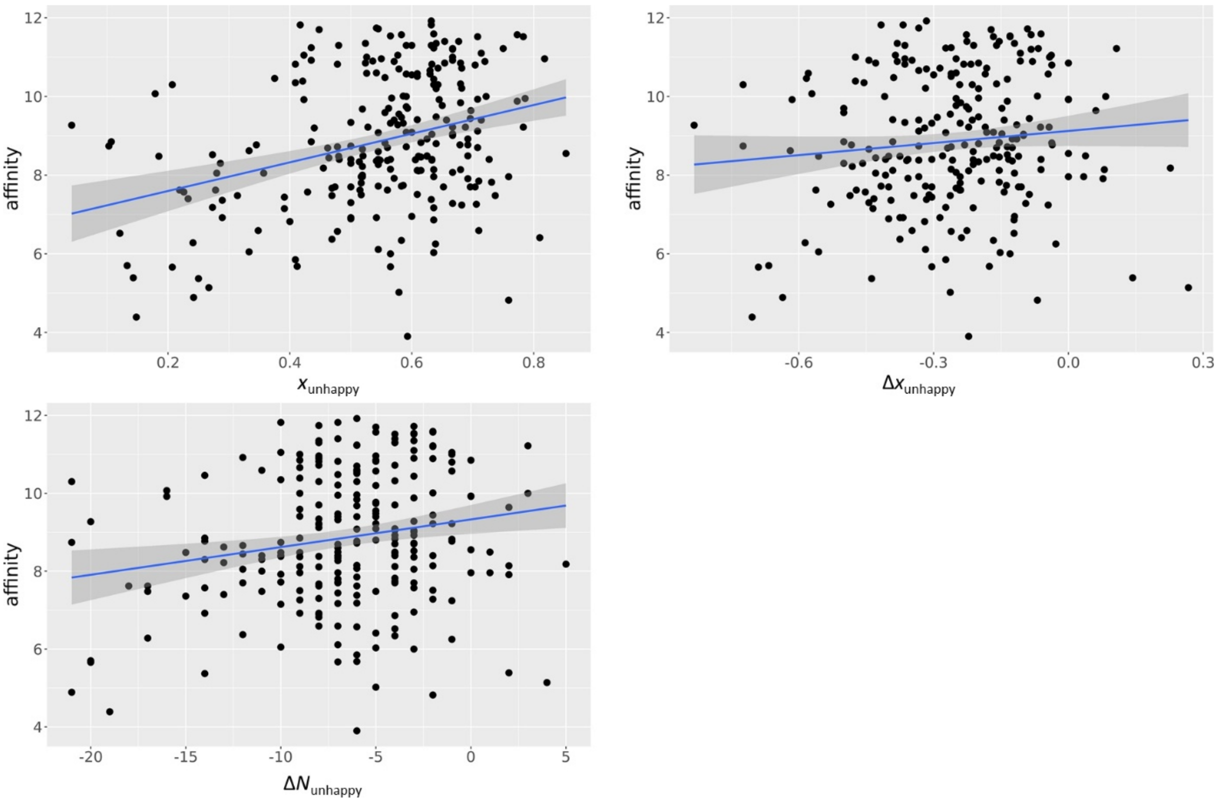

Figure 76: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for HIV1PR complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

HSP90

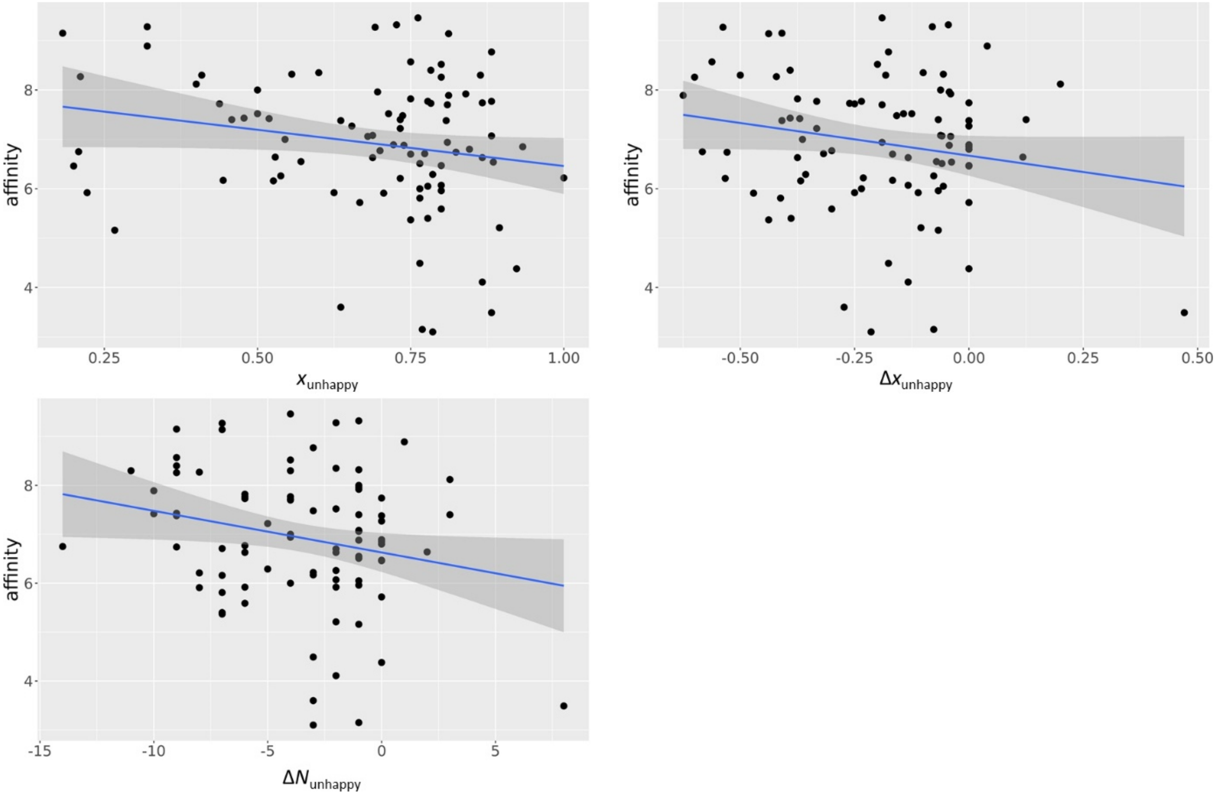

Figure 77: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for HSP90 complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

MMP12

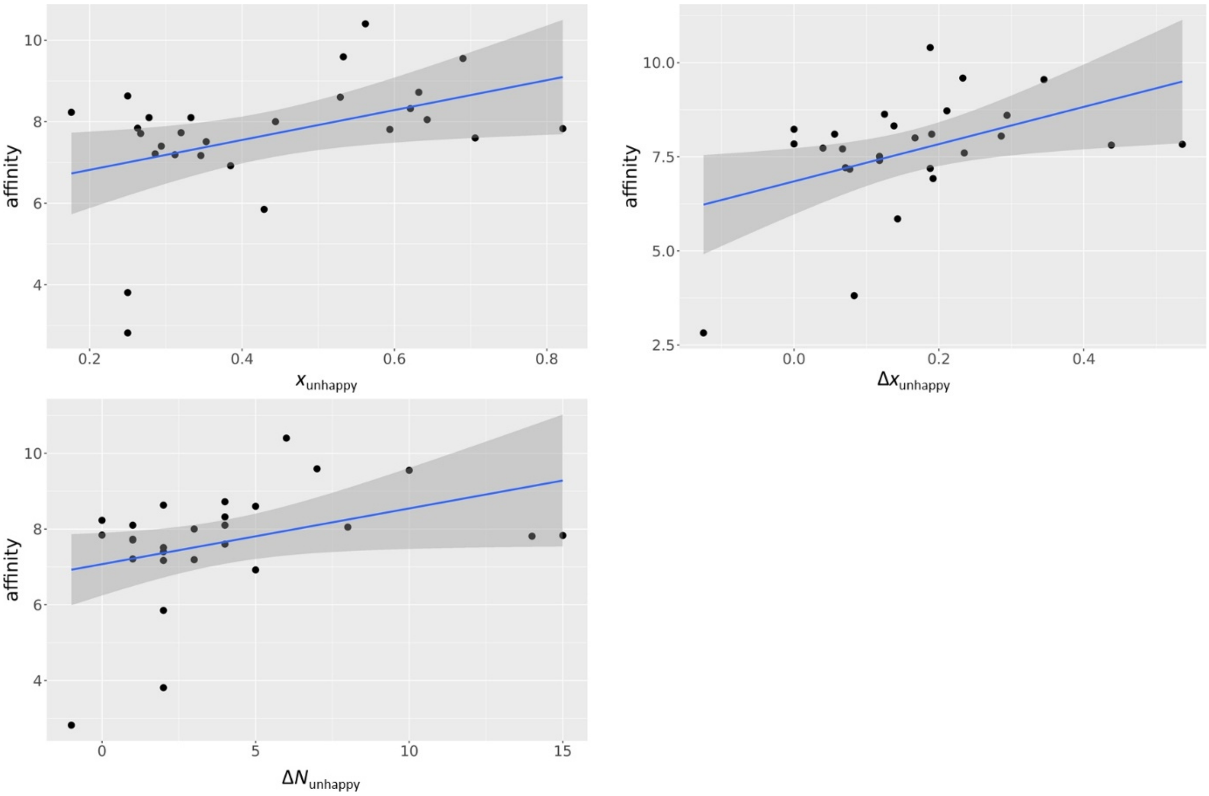

Figure 78: Scatterplot of ligand affinity (pK$_i$/pK$_d$) and the parameters x$_{unhappy}$, Δx$_{unhappy}$, and ΔN$_{unhappy}$ for MMP12 complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

NA

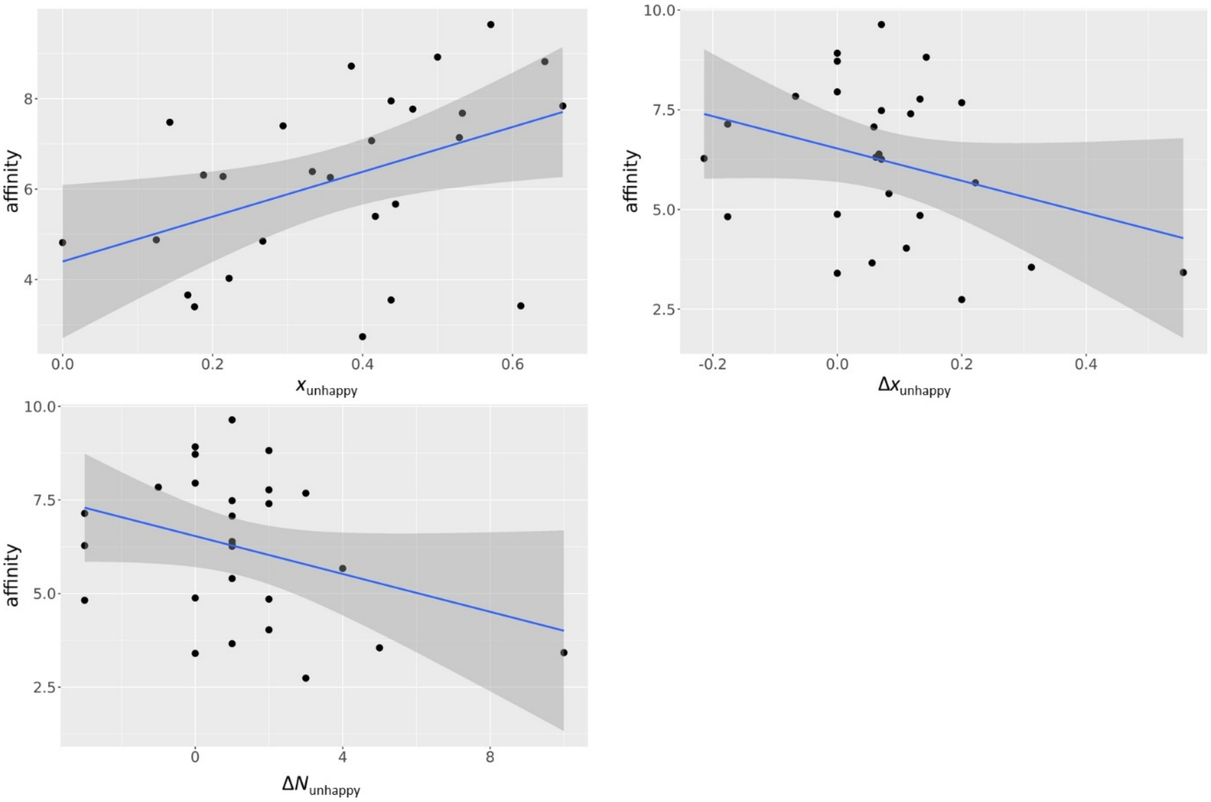

Figure 79: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for NA complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

Thermolysin

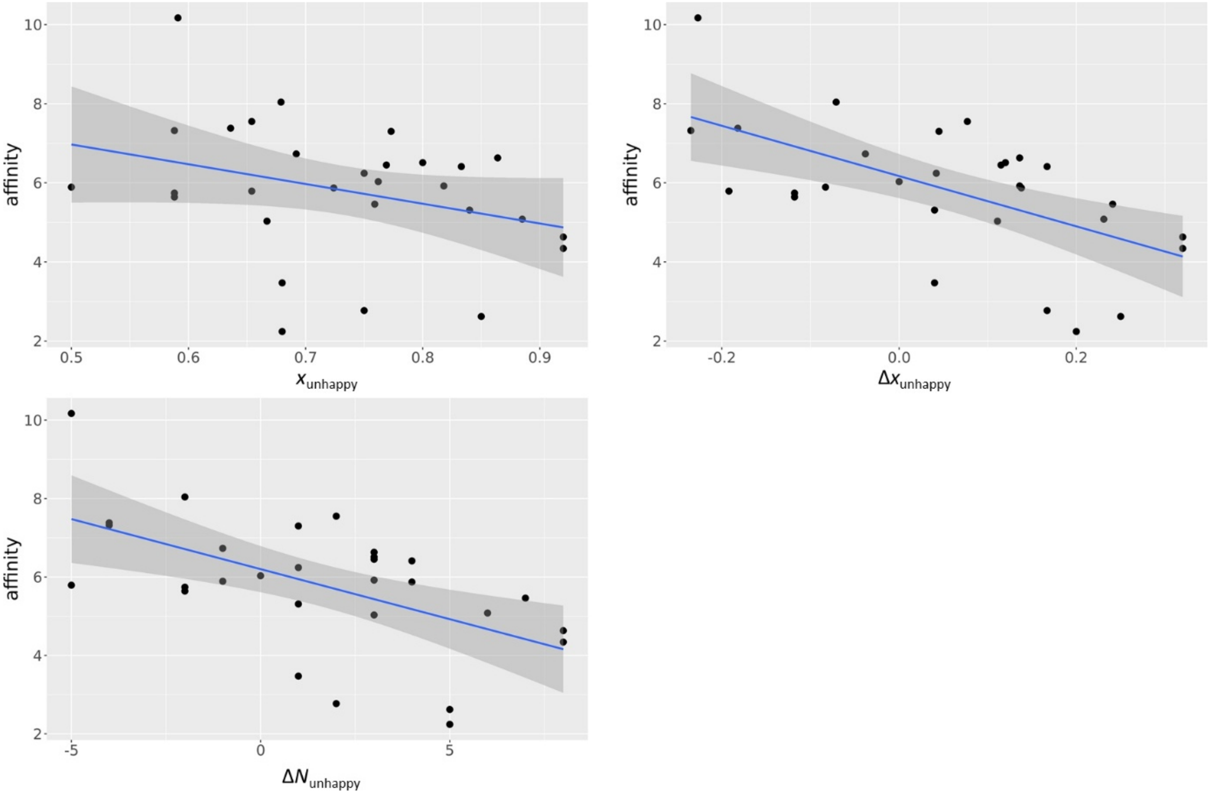

Figure 80: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for thermolysin complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

Thrombin

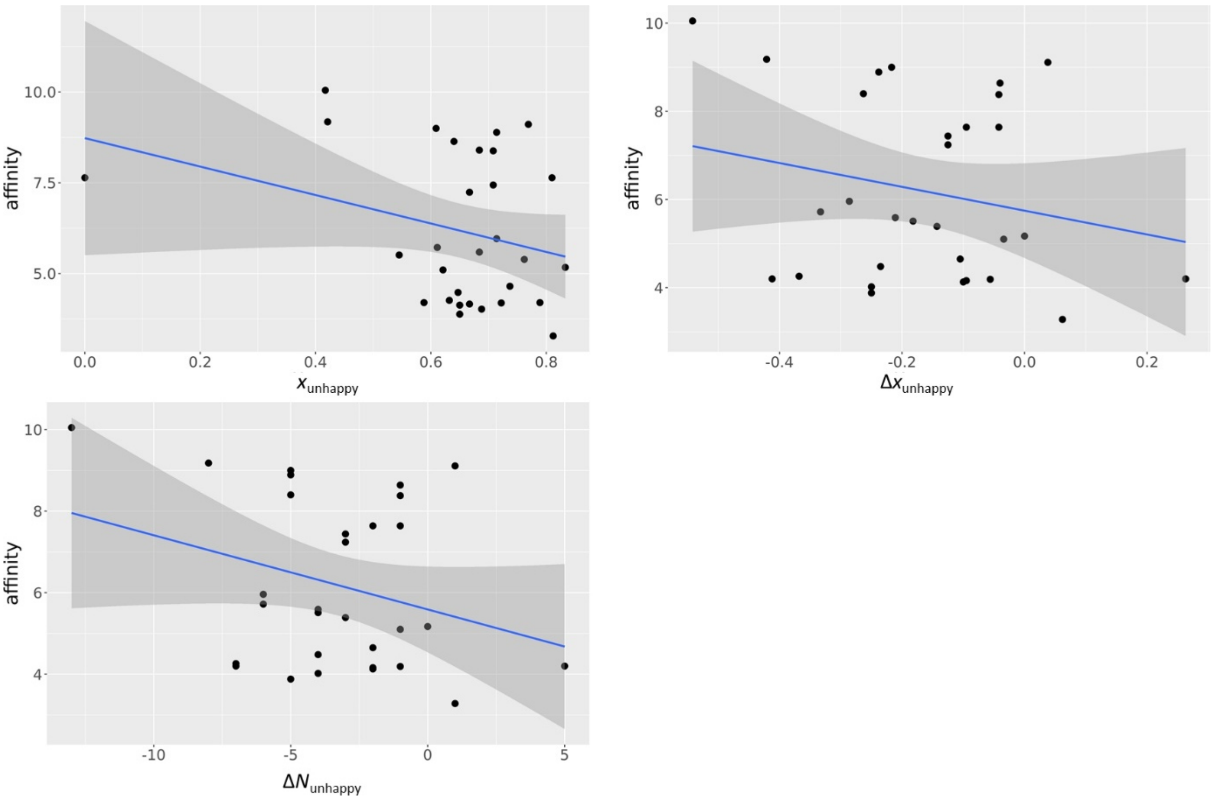

Figure 81: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for thrombin complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

Trypsin

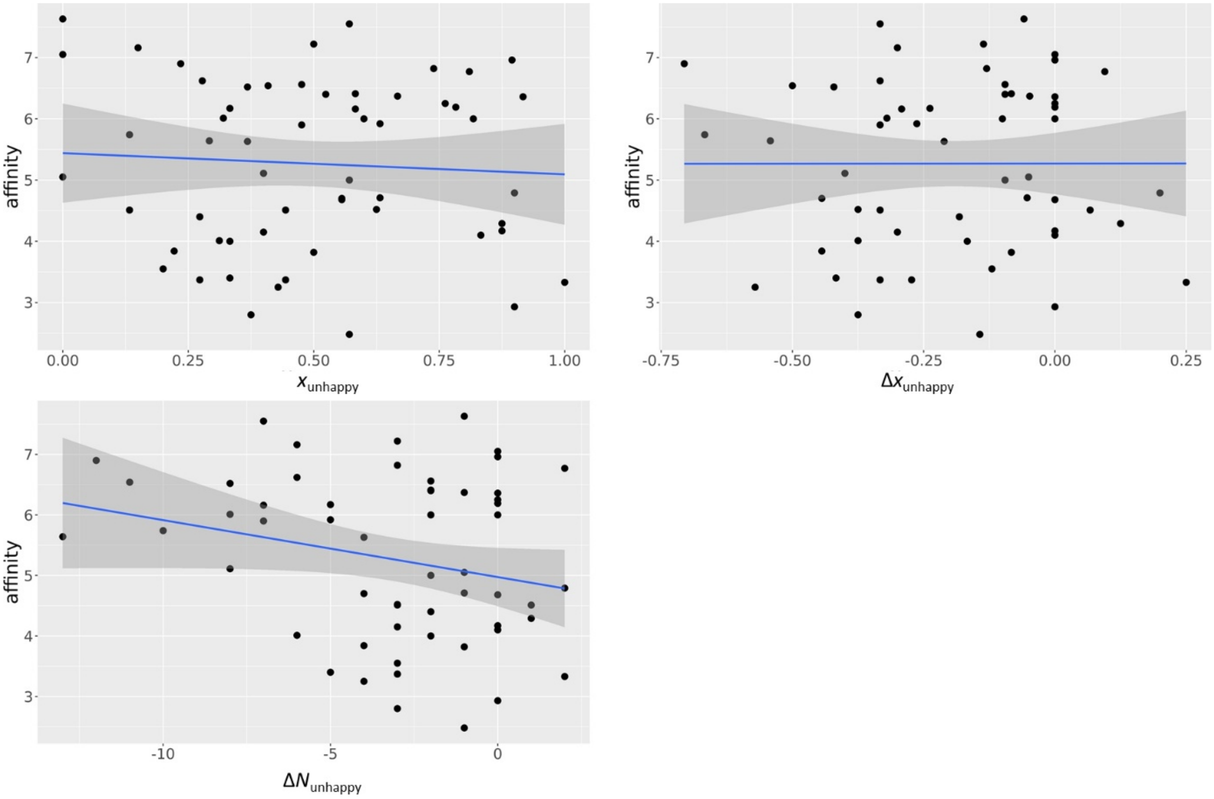

Figure 82: Scatterplot of ligand affinity (pK_i/pK_d) and the parameters $x_{unhappy}$, $\Delta x_{unhappy}$, and $\Delta N_{unhappy}$ for trypsin complexes in the used PDBbind refined subset with linear regression (blue) and respective 0.95 level of confidence interval (grey) as calculated by ggplot2 functionalities in R.

7.6 Ligand-probe matchscore vs. ligand affinity (PDBbind core set 2013)

Table 27: Binding affinities and the calculated ligand-probe matchscore for the complex structures in the PDBbind core set 2013.

PDB	pK_{aff}	s
4djr	11.52	0.186
2j62	11.34	0.165
2x00	11.33	0.152
1mq6	11.15	0.272
1lor	11.06	0.411
3utu	10.92	0.217
4gid	10.77	0.306
3myg	10.7	0.200
4tmn	10.17	0.174
1igj	10	0.127
1e66	9.89	0.330
3pe2	9.76	0.289
3g0w	9.52	0.246
2yki	9.46	0.333
3fv1	9.3	0.315
1sqa	9.21	0.259
2xy9	9.19	0.133
3su3	9.13	0.145
3gnw	9.1	0.185
2p4y	9	0.021
3dd0	9	0.249
3nw9	9	0.296
3uri	9	0.207
3e93	8.85	0.355
2obf	8.85	0.347
2zcq	8.82	0.381
2wtv	8.74	0.313
1hfs	8.7	0.278
1jyq	8.7	0.293
2pcp	8.7	0.098
3ge7	8.7	0.354

PDB	pK_{aff}	s
3nox	8.66	0.260
3f17	8.63	0.308
3ejr	8.57	0.375
2fvd	8.52	0.278
2vw5	8.52	0.143
2xbv	8.43	0.306
2qbp	8.4	0.263
3b68	8.4	0.247
1h23	8.35	0.522
2v7a	8.3	0.285
2cbj	8.27	0.148
1nvq	8.25	0.312
3l3n	8.18	0.375
2pq9	8.11	0.331
3ag9	8.05	0.070
2cet	8.02	0.342
3cyx	8	0.235
1o3f	7.96	0.190
2x8z	7.96	0.452
4g8m	7.89	0.187
1u1b	7.8	0.273
2zwz	7.79	0.421
2g70	7.77	0.263
2d1o	7.7	0.379
2zjw	7.7	0.256
3f3e	7.7	0.417
2xb8	7.59	0.298
1f8c	7.4	0.385
3su2	7.35	0.150
3kv2	7.32	0.063
3pww	7.32	0.184
2vvn	7.3	0.345
1kel	7.28	0.359
2ole	7.25	0.156
3gcs	7.25	0.570
1oyt	7.24	0.263

PDB	pK_{aff}	*s*
2vot	7.14	0.267
1xd0	7.12	0.136
1z95	7.12	0.244
4dew	7	0.180
2d3u	6.92	0.268
3uex	6.92	0.241
3gbb	6.9	0.433
3oe5	6.88	0.196
2zcr	6.87	0.383
3s8o	6.85	0.177
2xnb	6.83	0.395
2xhm	6.8	0.264
2jdu	6.72	0.261
4djv	6.72	0.319
2iwx	6.68	0.159
1sln	6.64	0.221
2yfe	6.63	0.290
3jvs	6.54	0.272
2weg	6.5	0.302
1r5y	6.46	0.479
2j78	6.42	0.270
10gs	6.4	0.318
1lol	6.39	0.587
2qbr	6.33	0.363
3k5v	6.3	0.358
1w3l	6.28	0.124
3bfu	6.27	0.289
1hnn	6.24	0.270
3vh9	6.24	0.204
1yc1	6.17	0.284
3bkk	6.08	0.344
3owj	6.07	0.294
1os0	6.03	0.268
2vl4	6.01	0.151
3huc	5.99	0.281
1q8u	5.96	0.268

PDB	pK_{aff}	*s*
4de1	5.96	0.283
3ehy	5.85	0.227
4des	5.85	0.360
2y5h	5.79	0.060
1o5b	5.77	0.207
1n1m	5.7	0.660
2x97	5.66	0.146
2wca	5.6	0.185
3su5	5.58	0.164
1f8b	5.4	0.276
2jdm	5.4	0.208
1gpk	5.37	0.604
2qft	5.26	0.133
3ueu	5.24	0.250
1w4o	5.22	0.235
1zea	5.22	0.139
2zxd	5.22	0.255
3ov1	5.2	0.171
3zso	5.12	0.433
3gy4	5.1	0.190
2yge	5.06	0.053
2gss	4.94	0.244
1p1q	4.89	0.187
2vo5	4.89	0.126
3d4z	4.89	0.563
2brb	4.86	0.304
3bpc	4.8	0.154
1q8t	4.76	0.189
3acw	4.76	0.265
1vso	4.72	0.228
3mss	4.66	0.235
1u33	4.6	0.145
2x0y	4.6	0.264
2wbg	4.45	0.257
3pxf	4.43	0.496
3u9q	4.38	0.333

PDB	pK_{aff}	s
2jdy	4.37	0.280
1a30	4.3	0.272
1w3k	4.3	0.131
3ivg	4.3	0.289
2qmj	4.21	0.098
3b3w	4.19	0.244
3ozt	4.13	0.501
4de2	4.12	0.170
1n2v	4.08	0.269
3kwa	4.08	0.167
2w66	4.05	0.342
2hb1	3.8	0.401
3nq3	3.78	0.233
2r23	3.72	0.239
1loq	3.7	0.454
3n7a	3.7	0.143
2v00	3.66	0.204
1f8d	3.4	0.395
1bcu	3.28	0.103
3zsx	3.28	0.406
1lbk	3.18	0.335
2ymd	3.16	0.421
2xdl	3.1	0.325
3imc	2.96	0.519
4gqq	2.89	0.439
3udh	2.85	0.439
3lka	2.82	0.266
3fcq	2.77	0.100
3fk1	2.62	0.440
3kgp	2.57	0.125
3b3s	2.55	0.206
3mfv	2.52	0.262
3dxg	2.4	0.212
3g2z	2.36	0.000
1qi0	2.35	0.067
1ps3	2.28	0.364

PDB	pK_{aff}	s
1uto	2.27	0.167
3ao4	2.07	0.307

7.7 Virtual screening of XIAP – ROC curves

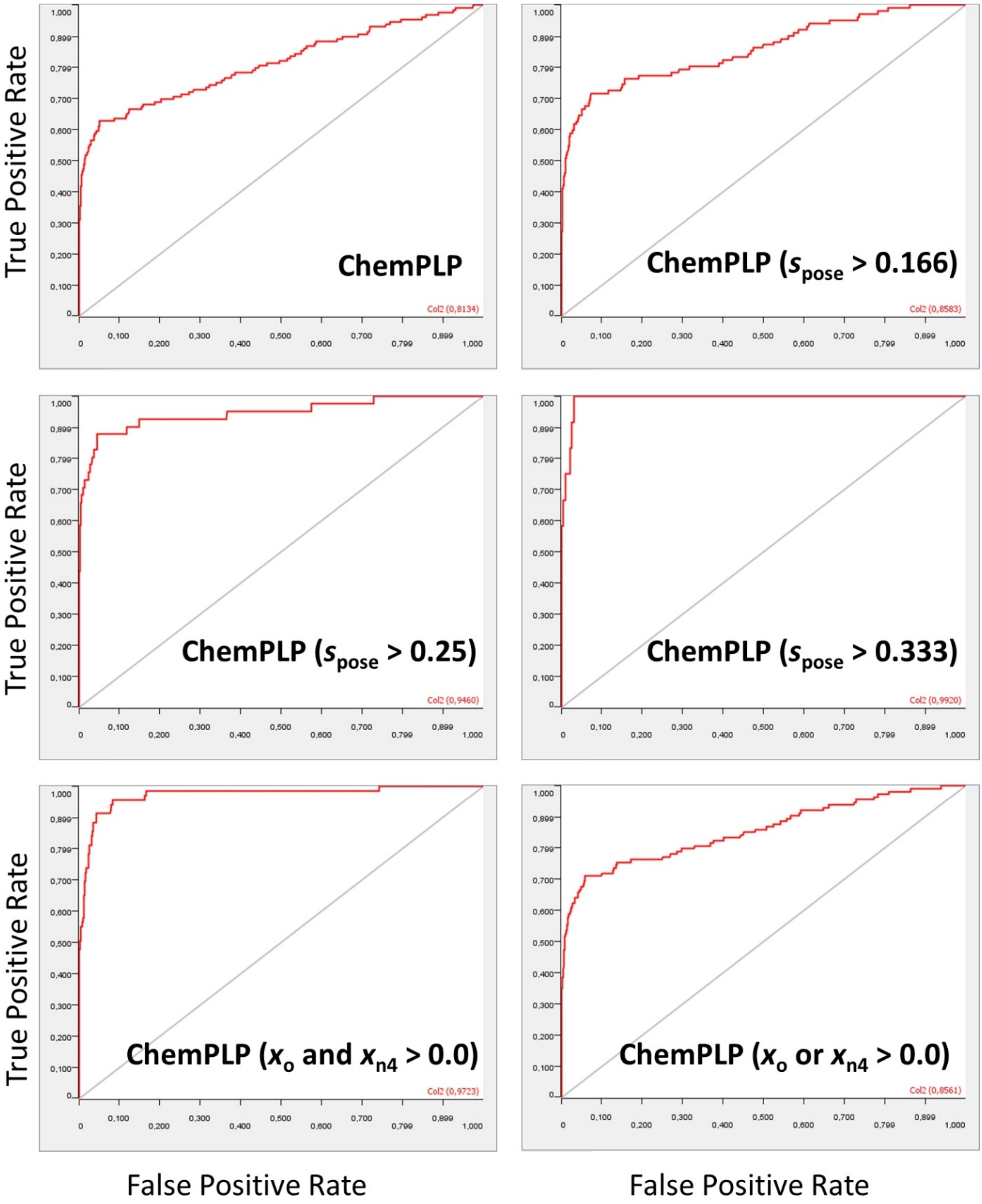

Figure 83: ROC curves for the scoring of poses obtained by docking of the DUD-E benchmark data set for XIAP by ChemPLP score alone and in combination with the ligand-probe matching score s_{pose} for filtering purposes; x_o and x_{n4} denote the matching with only o and n4 probe as defined in Eq. (77).

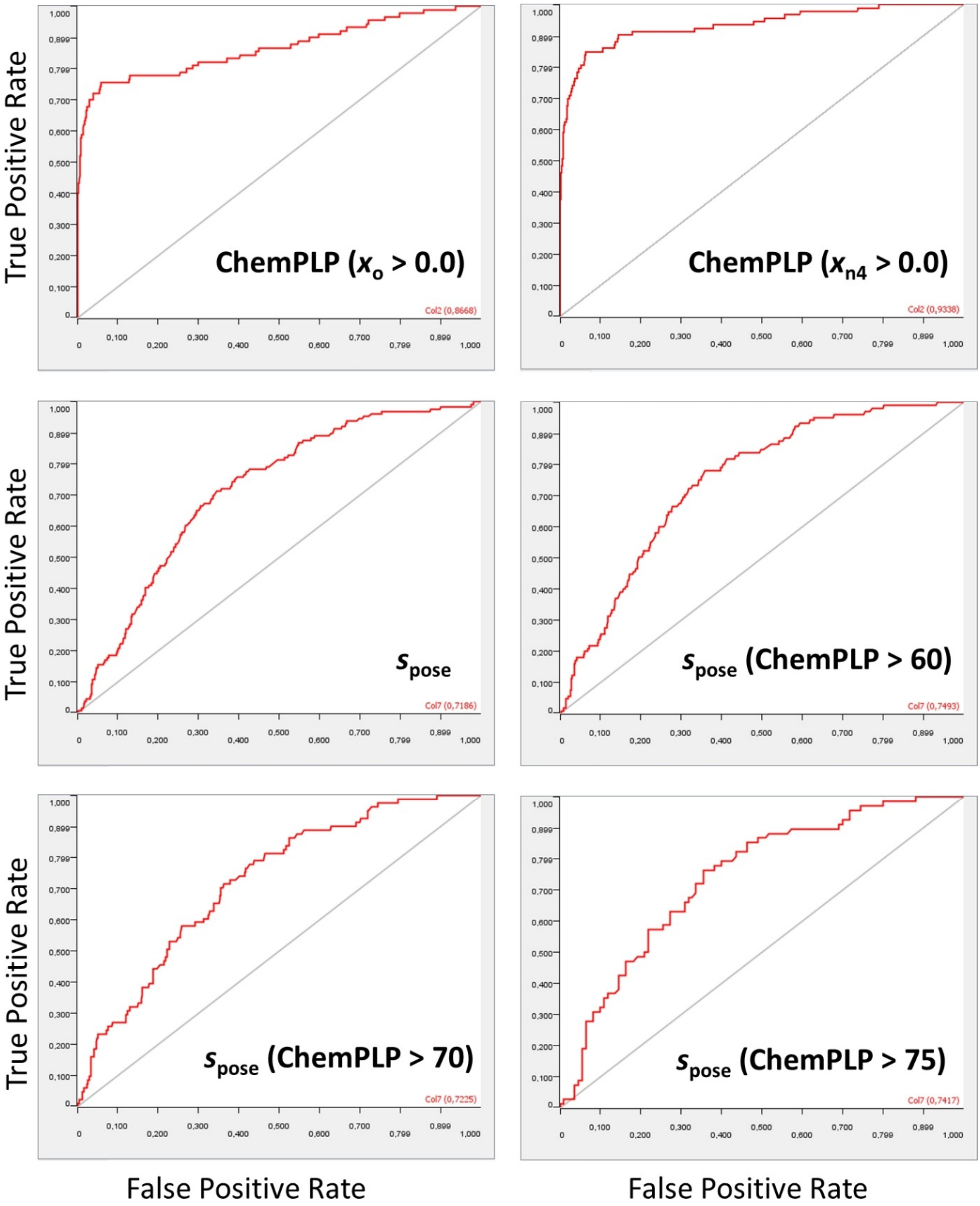

Figure 84: ROC curves for the scoring of poses obtained by docking of the DUD-E benchmark data set for XIAP by ChemPLP score, the ligand-probe matching score s_{pose} and different combinations of them for filtering purposes; x_o and x_{n4} denote the matching with only o and n4 probe as defined in Eq. (77).

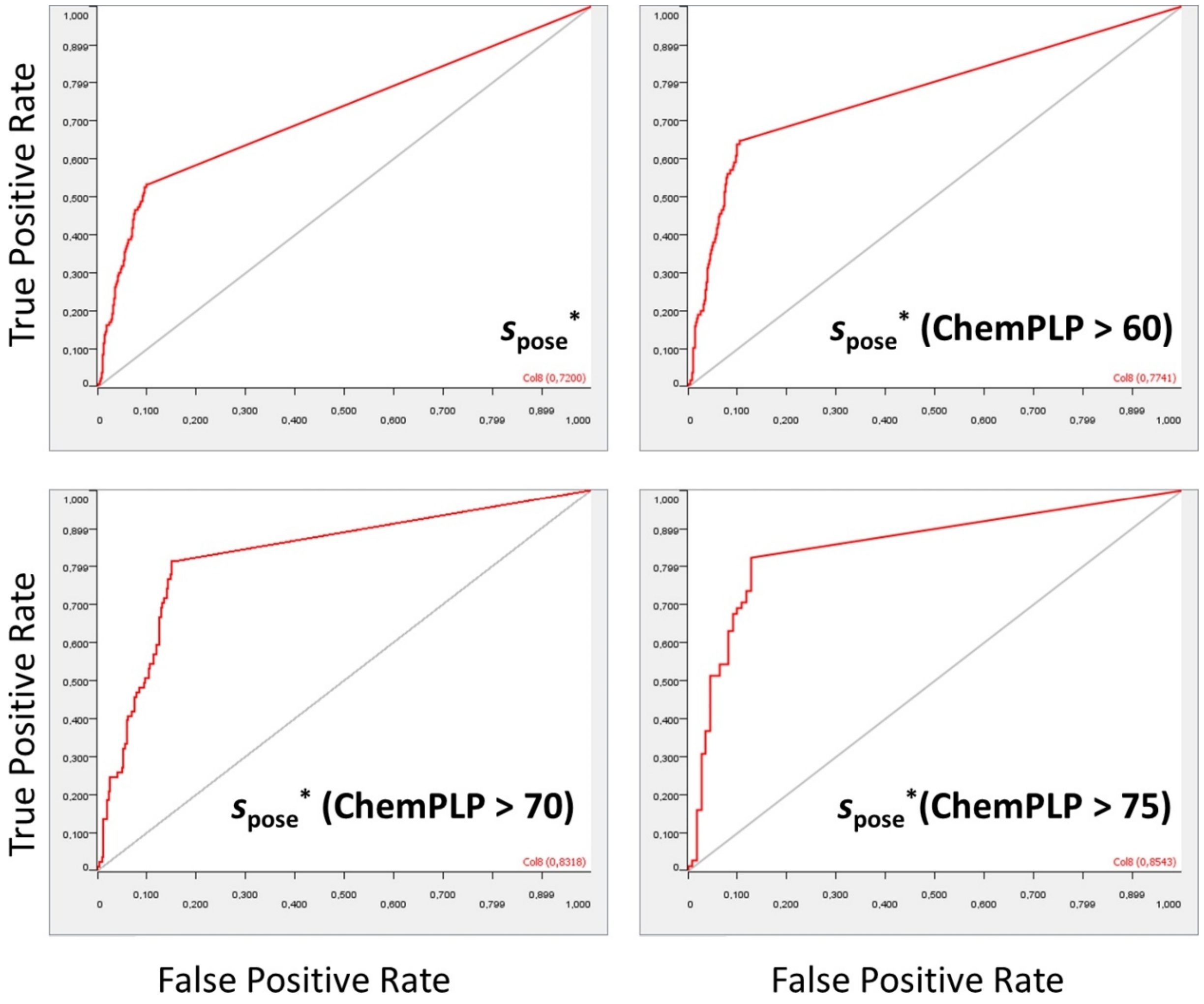

Figure 85: ROC curves for the scoring of poses obtained by docking of the DUD-E benchmark data set for XIAP by ChemPLP score, the ligand-probe matching score s_{pose}^* as defined in Eq. (79) and different combinations of them for filtering purposes.

7.8 TEAD4 scores

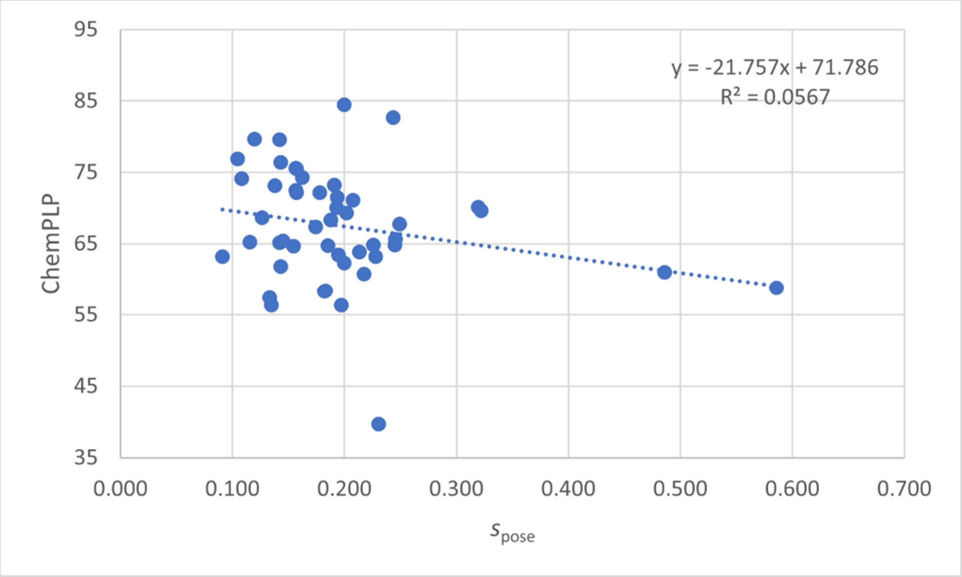

Figure 86: Scatter plot of ChemPLP scores and s_{pose} values as defined 4.2.2. for the docking of the data set provided by the Brunschweiger group in TEAD4 structure 6q36. The raw data can be found in the electronic appendix (Electronic_Appendix/ TEAD4/ Docking/).

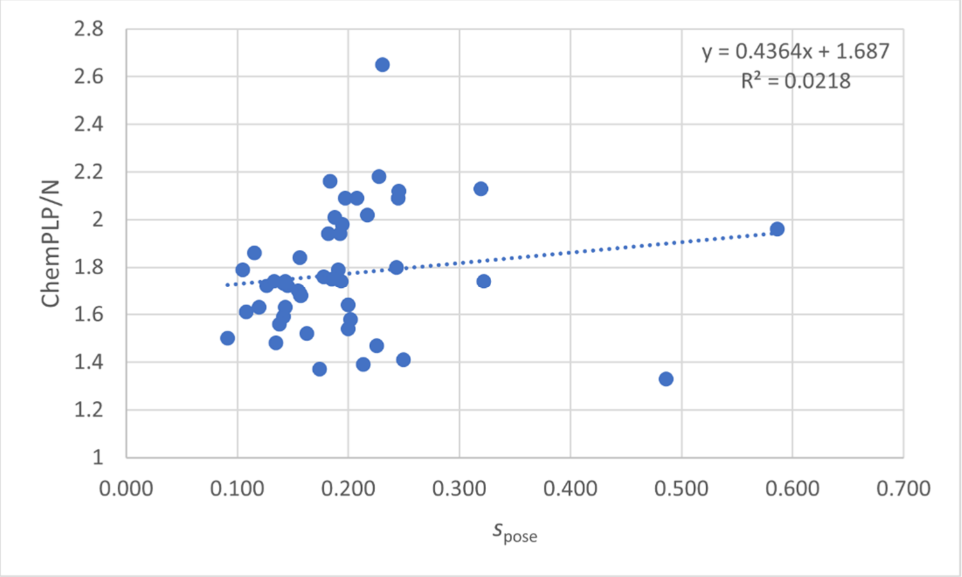

Figure 87: Scatter plot of normalized ChemPLP scores (score divided by number of heavy atoms) and s_{pose} values as defined 4.2.2. for the docking of the data set provided by the Brunschweiger group in TEAD4 structure 6q36. The raw data can be found in the electronic appendix (Electronic_Appendix/ TEAD4/ Docking/).