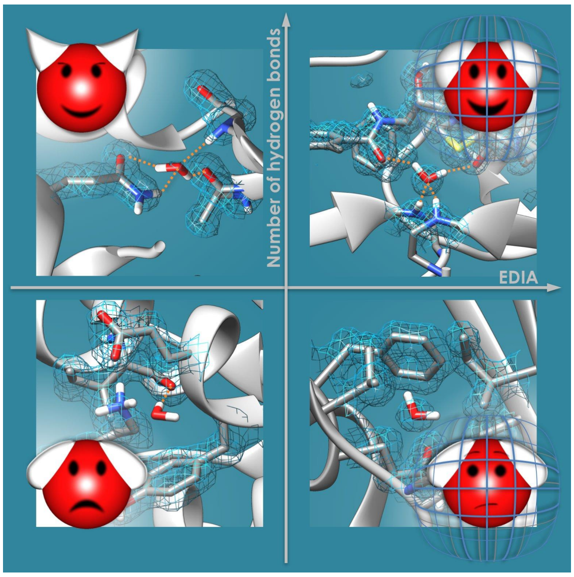
Water Molecules Within the HYDE Scoring Function:

Placement, Optimization, and Energetic Contributions



Cumulative Dissertation
with the aim of achieving the degree

Dr. rer. nat.

at the Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics
of Universität Hamburg

submitted by

Eva Nittinger

born in Wesel

Hamburg, March 2018

1st Reviewer: Prof. Dr. Matthias Rarey

2nd Reviewer: Prof. Dr. Johannes Kirchmair

3rd Reviewer: Prof. Dr. Peter Kolb

Date of thesis defense: 15.06.2018

Danksagung

An dieser Stelle möchte ich mich bei all denen bedanken, die mich während meiner Promotion unterstützt und begleitet haben.

Allen voran möchte ich mich bei meinem Betreuer Prof. Dr. Matthias Rarey für das interessante Promotionsthema, sein entgegebgebrachtes Vertrauen, seine Unterstützung und seine stets offene Tür für hilfreiche Diskussionen bedanken.

Ich bedanke mich auch bei Prof. Dr. Johannes Kirchmair und Prof. Dr. Peter Kolb für die Begutachtung meiner Dissertationsschrift.

Dr. Gudrun Lange und Dr. Robert Klein möchte ich für stetig neue Ideen, ausgiebige Diskussionen und die gute Zusammenarbeit sowohl während, als auch nach unserem Kooperationsprojekt danken. Ebenfalls möchte ich an dieser Stelle den Kooperationspartern der BioSolveIT danken, speziell während der Weiterentwicklung des HYDE Projekts.

Many thanks to Dr. Daniel F. Ortwine and Dr. Paul Gibbons for an interesting topic, a great supervision, and fruitful discussions during my internship.

Für die tollen Jahre, die Diskussionen sowohl fachlicher, als auch nicht fachlicher Themen bedanke ich mich bei allen Mitarbeitern der AMD-Arbeitsgruppe. Dr. Nadine Schneider möchte ich hierbei besonders für die Einführung zu Beginn meiner Promotion und ihren immer verfügbaren Ratschläge danken. Dr. Karen Schomburg danke ich füer ihre Unterstützung und die Zusammenarbeit während des HYDE Projekts und im weiteren Verlauf meiner Promotion. Bei Dr. Therese Inhester bedanke ich mich für die erfolgreiche Zusammenarbeit während der Entwicklung von *NAOMI*nova. Kai Sommer danke ich für seine Antworten auf alle erdenklichen Fragen, Dr. Stefan Bietz für die stets ausführliche Beantwortung von Fragen zur NAOMI Bibliothek, Thomas Otto und Florian Flachsenberg für ihr hilfreiches Wissen während der Softwareentwicklung und Rainer Fährrolfes für die Bereitstellung des Review-Servers für *NAOMI*nova. Agnes Meyder möechte ich für die Zusammenarbeit während des HYDE Projekts und dessen Entwicklung danken. Melanie Geringhoff danke ich für die Hilfe mit jeglichen kleineren und größeren Problemen.

Für die Finanzierung des Projekts möchte ich mich bei dem Bundesministerium für Bildung und Forschung, dem Cluster Biokatalyse2021, Bayer CropScience, sowie der BioSolveIT bedanken.

Ganz besonders möchte ich mich bei meinen Eltern für die Ermöglichung meines Studiums und die immerwährende Unterstützung und ihre nie versiegenden Ratschläge während meiner Promotion bedanken.

Für das Korrekturlesen möchte ich mich besonders bei Niklas Nittinger bedanken. Zudem bin ich ihm für seine nie endende Motivation, die unendliche Geduld und fortwährende Unterstützung in jeglicher Situation unbeschreiblich dankbar.

Kurzfassung

Wassermoleküle übernehmen im biologischen Kontext verschiedene wichtige Rollen – von einem reinen Lösemittel, über strukturelle Eigenschaften wie der Vermittlung von Wasserstoffbrücken, hin zu katalytischen Funktionen. In den letzten Jahren wurde daher immer mehr die Rolle der Wassermoleküle in der Entwicklung von Medikamenten, Herbiziden und biokatalytischen Fragestellungen miteinbezogen. Aus diesem Grund wurden über die letzten Jahre diverse Methoden für die Analyse von Wassermolekülen entwickelt – von der strukturellen Analyse von kristallografisch aufgelösten Wassermolekülen, bis zur Vorhersage von Wassermolekülen in biologischen Strukturen und deren energetischen Beiträgen. Trotz der Menge und Varianz an verfügbaren Methoden bleiben diverse Fragen, insbesondere der energetische Beitrag von Wassermoleküle, ungeklärt. Hinzu kommt, dass viele der gebräuchlichen Methoden rechenintensiv sind und nur auf wenigen, selektierten Daten evaluiert wurden.

In der hier präsentierten Arbeit wird eine Methode für die Platzierung von Wassermolekülen mit anschließender Bewertung auf Basis der zuvor entwickelten HYDE Bewertungsfunktion dargestellt. Um Wassermoleküle korrekt zu platzieren, muss der zur Verfügung stehende Raum in Protienstrukturen richtig erkannt werden. Basierend auf einer Analyse von Proteinstrukturdaten wurden Wasserstoffbrückengeometrien definiert, die im Folgenden genutzt wurden um potenzielle Wasserpositionen zu definieren. Die Identifikation des so genannten freien Raums wurde im Anschluss genutzt, um explizit Wassermoleküle zu platzieren. Diese Wasserpositionen wurden im Anschluss mit HYDE bewertet, um deren energetische Beiträge abzuschätzen.

Neben der Methodenentwicklung wurde besonderer Wert auf dessen Evaluierung gelegt. Insbesondere Wassermoleküle stellen eine Herausforderung dar, da experimentelle Ergebnisse für einzelne Wassermoleküle nur schwer zu erhalten sind. Die einzige experimentell verfügbare Quelle, die in ausreichender Qualität und Quantität für Wassermoleküle verfügbar ist, ist die Elektronendichte, die bei der Proteinstrukturaufklärung entsteht. Auf Basis der Elektronendichte wurde das Maß EDIA (*Electron Density of Individual Atoms*) entwickelt, um automatisiert Kristallwasser mit ihrer experimentellen Grundlage abzugleichen. Mit Hilfe des EDIA wurde ein hochaufgelöster Datensatz zusammengestellt, welcher im Verlauf der Methodenentwicklung genutzt wurde, um die entwickelte Platzierungsstrategie und die Bewertung der Wassermoleküle zu validieren. Die Platzierung erreicht eine hohe Sensitivität, wobei 80% der vorhergesagten Wasser in einem Abstand von weniger als 1.0 Å zu kristallographischen Wassermolekülen platziert werden. Zusätzlich zu der konsistenten Platzierung von Wassermolekülen, erziehlt die Methode eine kurze Laufzeit, was im Folgenden die Analyse und Bewertung von Protein-Ligand-Komplexen sowie den Einfluss von Proteinflexibilität auf umgebende Wassernetzwerke ermöglicht.

Abstract

Water molecules play an important role in biological complexes, from simply surrounding solvent, to structural aspects such as the mediation of hydrogen bonds, to catalytic functionalities. Over the last years, many different methods and programs were developed for the analysis of water. These range from analyzing water molecules observed in crystallographically resolved protein structures, to predicting potential water positions and evaluating their energetic contribution to the binding affinity. However, despite the large amount of available tools, many questions, such as a the energetic contribution of water molecules, still remain a challenge. In addition, most frequently applied software solutions require substantial computational resources and are validated on small amounts of selected protein complexes.

The presented work shows a strategy for placement and subsequent energetic evaluation of water molecules based on the previous developed HYDE scoring function. First, potentially available space for a water molecule inside a protein structure has to be recognized correctly. Therefore, hydrogen bond geometries were analyzed within a large set of protein structures. Based on the derived hydrogen bond geometries, suitable positions within the protein structure were defined. This available space was used to place water molecules explicitly and, finally, score those positions with HYDE.

Beside the method development, great effort was put on its evaluation. Especially the validation of water molecules is a challenging task and no large-scale data set existed. Since single water molecule are very difficult to measure experimentally, adequate data has to be compiled for validation purposes. The electron density, the only experimental evidence sufficiently available for water molecules, was exploited. Using the electron density, a metric for the automatic evaluation of water positions with their underlying electron density was developed – EDIA (Electron Density of Individual Atoms). This way, a high-resolution data set with well resolved water molecules was compiled. This large-scale data set was used for the validation of the water placement and scoring procedure. The water placement strategy achieves a high sensitivity of 80% for placing water molecules within 1.0 Å distance to a crystallographically observed one. Due to a short run-time, the water placement procedure displays a solid foundation for further evaluation of protein-ligand complexes, the effects of protein flexibility on the surrounding water network, or scoring of protein-ligand interfaces with a consistent representation of water molecules.

Contents

1
Introduction

Three major components contribute to the composition of cells – water, inorganic ions, and organic molecules. Over 70% of the cellular mass is water.[1] This already shows the importance of water molecules in nature. In this thesis, the relevance of water molecules for proteins will be discussed. They not only play a passive role as surrounding solvent, but also fulfill active roles such as mediating interactions within protein structures[2–5] or being part of a reaction catalyzed by a protein.[6–8]

Therefore, water molecules are of interest in pharmaceutical, biotechnological as well as agricultural research. Throughout the last years, many different developments in the field of chemoinformatics have helped to guide the research in these areas in a more rational way. Time consuming wet-lab experiments as well as the work load could be reduced.

A great improvement in structure-based design is the constant enhancement of crystallography methods to elucidate protein structures with high resolution.[9] This allows more detailed insights into the structural composition of proteins, more precise evaluations and, by this, also a more detailed analysis of water molecules. Especially for water molecules high-resolution structures are mandatory to enable the crystallographer to correctly model the water molecules into the protein structure. In order to resolve water molecules, a resolution of at least 2.7 Å is needed.[10] However, water positions in structures with resolution lower than 1.8 Å are less reliable than water molecules in better resolved structures.[11] Additionally, in order to observe a continuous hydration layer at the protein surface, a resolution of better than 1.6 Å is needed.[12]

Due to the different roles water molecules can fulfill, they are relevant in many chemoinformatics methods from structure-based virtual screening (SBVS), to docking, to scoring functions. An analysis of the PDBbind 2007 data set[13] revealed water mediated interactions in 96% of protein-ligand complexes.[14] This exemplifies the need of a correct representation of water molecules within these methods.

Previously, the scoring function HYDE[15–19] was developed for the estimation of protein-ligand binding affinities. HYDE is a physics-based approach relying on **HY**drogen bonding and **DE**yhdration. Herein, water molecules display an essential part, but are currently not handled adequately. Two aspects are of major interest. First, a consistent treatment of water molecules within the HYDE scoring function. Second, a consistent representation of water molecules in protein structures, which is necessary to realize point one. The resolution of the protein structure may not be good enough to resolve water molecules or the structure has not been crystallized, i.e. protein-ligand docking poses, which means that water molecules are not available. A consistent integration and availability of water molecules within protein structures, especially within protein-ligand interfaces, is necessary.

In this thesis, a method for placing water molecules into the structural model of proteins is described. Beginning with the structural evaluation of water molecules, to the correct identification of available space for water molecules within protein structures, to, finally, the placement of structurally relevant water molecules. Subsequently this method is combined with the HYDE scoring function. Herein, the estimation of individual water scores is examined. This is of special interest for drug design strategies, i.e. to predict which water molecules would be best to target with a ligand in order to gain affinity.

Relevance of Water Molecules

'Water molecules appear to be both
the cement that fills crevices between amino acid building blocks,
and the lubricant that allows motion of these building blocks.' [*]

In the following subsections, different aspects of water molecules are described – from available experimental data for water molecules, to structural aspects, thermodynamic characteristics, available software solutions for placing and scoring water molecules, and the relevance of water molecules in chemoinformatic methods.

[*] Levitt, M. & Park, B. H. Water: now you see it, now you don't. Structure 1, 223–226 (1993).

1.1 Water Molecules in Protein Structures

Many aspects around water molecules are of interest in the area of chemoinformatics. Beginning with the availability of experimental data for water molecules to allow their correct representation. Further, structural characteristics found in protein structures are important for understanding the role of water molecules. Last, the thermodynamics displayed by a water and the influence of thermodynamic effects by water molecules is of great relevance for method development, but also to comprehend the underlying mechanisms of, i.e., protein-ligand binding events.

1.1.1 Experimental Data and Validation

Water molecules are difficult to examine experimentally. Especially when it comes to their energetic contribution, problems arise. Due to the fact that a water molecule in a protein-ligand interface is displaced by extending a ligand, the difference in measured binding affinity is always the combination of water displacement together with the ligand extension.[20–22] Therefore, other means of validation have to be identified for the evaluation of water molecules.

The largest resource of experimental data are protein structures solved by X-ray crystallography (Figure 1.1). Those structures are readily accessible from the Brookhaven Protein Data Bank (PDB).[23] Thanks to the increasing focus on data quality within the last years, structure factors are nowadays mandatory for newly resolved structures deposited in the PDB. With a constantly growing number of available structures, quantitatively as well as qualitatively, the PDB is an ideal data resource.

Already the protein structure displays a model of the experimentally collected data, the diffraction patterns. Different metrics exist to validate the model with its underlying data. Accessible directly from the PDB file are occupancy and temperature factor (B factor). The occupancy is given for every atom and contains information about alternate locations. This is, however, not always correctly used and especially for water molecules difficult to interpret. The B factor contains information about the local motion of the structure. Since the B factor is dependent on the refinement procedure[24–26] it can be artifactual, especially if crystal contacts are available but not considered adequately. Furthermore, the B factor does not contain information about an atom being resolved by electron density, but about its structural flexibility as well as disorder. Other metrics exist that consider the underlying electron density data: real-space R factor (RSR),[27] normalized RSR (RSR-Z),[28] real-space R correlation coefficient (RSCC),[29] real-space difference density Z score (RSZD), and real-space observed density Z score (RSZO).[30] However, all of them have strengths as well as weaknesses. For example the often used RSCC exhibits problems especially for water molecules. This is due to atoms with weak densities but correct intensity distributions. Thus, even a low resolution can result in a good score. For more detailed information on the different metrics see D2 and D3.

Apart from X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy displays a resource for protein structural data. Further, electron microscopy (EM) has evolved dramatically throughout the last years. From previously only low-resolution structures, high-resolution EM structures are emerging and will become more and more available throughout the next years.[9] However, of the overall available structures, NMR and EM structures display only a minority (Figure 1.1).

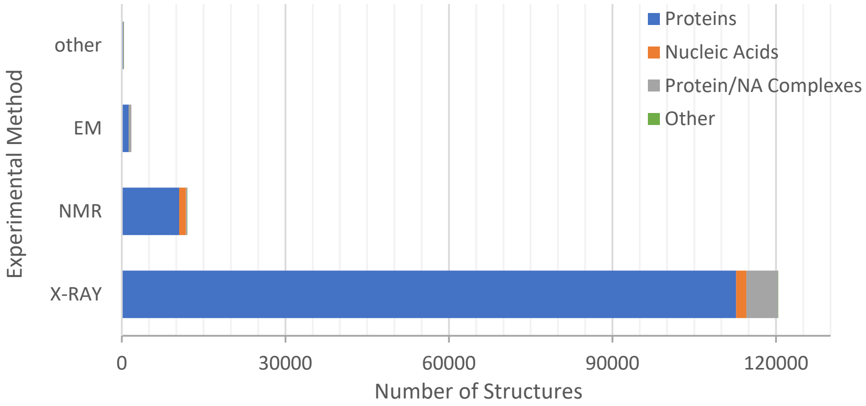


Figure 1.1: Number of available structures in the PDB[23] differentiated by experimental method and composition (as of October 2017).

One critical aspect about X-ray structures are the experimental conditions, which are commonly used to obtain the diffraction patterns. Usually, diffraction data are collected at cryogenic temperatures (100 K). Generally, it is assumed that the cooling process disturbs the natural structure of the macromolecule only minimally. Due to this process, however, only one snapshot of the naturally fluctuating macromolecular structure is captured. Apart from the exact location of water molecules, the different properties between 'hydration water' and bulk water were subject of analysis. Different terms used for hydration aspects are explained in Figure 1.2.

Studies have been conducted analyzing the differences between cryogenic and room temperature X-ray structures as well as X-ray structures versus NMR structures.[31–33] Results showed that cryocooling leads to remodeling of conformational distributions of proteins as well as to over-packed, smaller structural models. Room temperature structures on the other hand can reveal protein motions necessary for its function or ligand binding. NMR spectroscopy has also been used to study the differences of the first hydration shell of DNA between solution and crystal.[34] This study concluded that at low temperatures the rotation of water molecules is less temperature dependent, which they attribute to constraints from the protein. Nuclear magnetic relaxation dispersion (NMRD) experiments were conducted to analyze hydration layer waters.[35] Compared to bulk water, water molecules in the hydration layer of the protein are roughly two times slower, with the exception of those trapped in surface cavities. Nakasako compared the number of water molecules in cryogenic structures with that in solution.[36] He concluded that based on small angle X-ray scattering (SAXS) measurements the amount of hydration is comparable between cryogenic structures and solution.

X-ray structures, moreover those elucidated at cryogenic temperatures, display the majority of available structures. In combination with the underlying electron density they are of great relevance for method development and validation. Overall, one has to bear in mind that cryogenic protein structures are only one snapshot of the structure in nature.

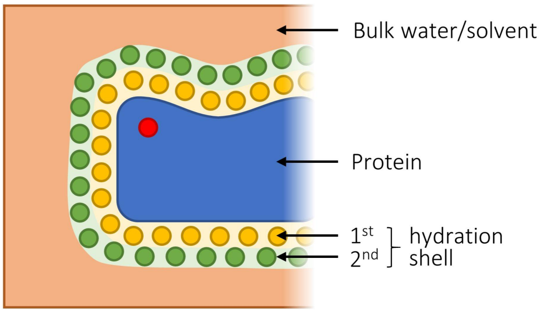


Figure 1.2: Hydration of macromolecules; blue: macromolecule, e.g. protein or DNA; red: buried water molecule; yellow: first hydration shell = water molecules directly surrounding the macromolecule typically within 3.5 Å with their properties affected by the macromolecule; green: second hydration shell = water molecules surrounding the macromolecule with their properties affected by the macromolecule; orange: bulk water/solvent = water molecules distant to macromolecular structures only in contact with other water molecules. For more information on protein hydration refer to a recent publication of Carugo.[37]

1.1.2 Structural Characteristics

Water molecules have different roles in and around protein structures and thus exhibit specific structural characteristics. NMRD experiments have shown that water molecules in protein structures, regardless of their position in a protein-ligand complex, are not fixed but in constant dynamic exchange with bulk waters.[38] Herein, water molecules in the protein vicinity are slower than bulk waters.[39] Elastic incoherent neutron scattering (EINS) displays another means of analyzing water molecules. Combet and Zanotti have identified interfacial waters as the driving force of local and large-scale motions in proteins.[40]

Water molecules contribute to several processes in protein structures: from the big picture, such as protein folding[41–44] or active/inactive transformations,[45] to small details, such as distance regulation,[46] proton transfer,[46–49] as well as energy storage.[50] Herein, flexibility aspects are a result of rather weakly bound water molecules. They still remain mobile and thus allow structural changes. Functionality on the other hand is due to tightly bound water molecules that are an integral part of the protein structure. This aspect is further emphasized in hyperthermophilic proteins. Molecular dynamics simulations were used to study differences of internal water molecules in hyperthermophilic and mesophilic proteins. Herein, hyperthermophilic proteins offered a slightly more favorable environment for water molecules compared to non-hyperthermophilic proteins.[51,52]

Apart from their overall roles, water molecules have been analyzed within proteins, within protein-ligand interfaces, as well as in protein-protein interfaces.[53–56] Water molecules bridging between protein and ligand often have three or more interactions compared to an average of one interaction in protein-protein interfaces.[57,58] Also within internal protein cavities, they form on average three hydrogen bonds with half the cavity waters occurring in clusters of two or more.[59] NMR studies showed that hydrophobic cavities, which often appear to be unsaturated in crystal structures, can contain mobile water molecules.[60] Using Monte-Carlo simulations, the smallest stable water cluster in hydrophobic binding sites was a trimer with three hydrogen bonds.[61] The implications of active site surface characteristics on water molecules has recently been analyzed by Haider *et al.*.[62] They used explicit solvent molecular dynamics (MD) simulations to study the reorganization of water molecules. Compared to bulk waters, they found favorable active site water molecules – with either enhanced water-water interactions or strong hydrogen bonds to the protein – as well as unfavorable water molecules. The latter ones result from surface constraints that hinder water molecules in adopting favorable orientations. Further, water molecules in protein-protein interfaces can be differentiated based on their interaction characteristics, bridging, i.e. interactions with both proteins, favorable for one protein, or no interaction with any protein.[63] Interestingly, 69% of water molecules of the last category (in total 18%) are in a hydrophobic environment, so called 'hydrophobic bubbles' (Figure 1.3).

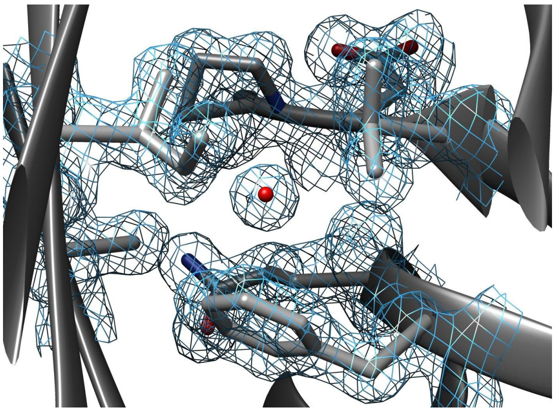


Figure 1.3: Hydrophobic enclosed water molecule H_2O-A-535 in Sec12 a component of the COPII vesicle budding machinery responsible for vesicle transport of proteins and lipids from the endoplasmic reticulum to the Golgi (PDBid 4h5i[64]); Mesh: electron density displayed at 1σ (Molecular graphics were created using UCSF Chimera[65]).

Another aspect that has been evaluated, are the binding partner preferences of water molecules. Indicated by the number of interactions, water molecules prefer hydrogen bonds to the protein backbone rather than to side chains.[58,63,66]

1.1.3 Thermodynamic Characteristics

The full picture, of how water molecules exactly contribute to binding affinity of protein-ligand as well as protein-protein complexes or in combination with DNA, has not been fully understood.[67,68] The following paragraphs summarize thermodynamic aspects in which water molecules have been analyzed and shown to contribute.

Estimation for the energetic contributions of water molecules have been made. Dunitz concluded from energetics of anhydrous salts and their corresponding hydrates that the upper bound for entropic cost is about 2 kcal mol^{-1} at 298K.[69] Ladbury on the hand estimated the upper enthalpic gain as -3.8 kcal mol^{-1}.[67] A combination of both estimates results in a free energy gain of -1.8 kcal mol^{-1} for transferring a water molecule from bulk into the protein or the active site.[70] This is in accordance with approximations from Cooper, who estimated the free energy range between -0.7 to -2.2 kcal mol^{-1}, depending on the number of hydrogen bond interactions the water molecule can participate in.[71]

Experimentally, the biophysical technique isothermal titration calorimetry (ITC) is applied to measure the thermodynamic signature of binding.[72,73] Herein, enthalpy changes (ΔH), association constant (K_a), and the stoichiometry of binding are derived. Those can be used to calculate the free energy of binding, also termed Gibbs free energy, (ΔG) as well as the entropy change (ΔS) using the following formula:

$$\Delta G = -RT \ln K_a = \Delta H - T\Delta S \tag{1.1}$$

R is the gas constant and T the temperature. ITC measurements in combination with structural and molecular dynamics simulations have shown in Abl-Src homology 3 domain (SH3) with the high affinity peptide p41 that interfacial water molecules need to be included to fully understand the binding process.[74] Herein, hydrophobic interactions in the protein binding site were complemented by a water-mediated hydrogen bond network of circumferential residues. For the application of ITC for diverse systems, from protein-ligand interactions, protein-peptide interactions, protein-protein interactions, to enzyme activity and kinetics, please refer to Ghai *et al.*.[75] Kimmer *et al.* have reviewed the influences of different parameters on ITC measurement, such as buffer or experimental set-up, to get a better estimate on the reliability and comparability of the generated data.[76]

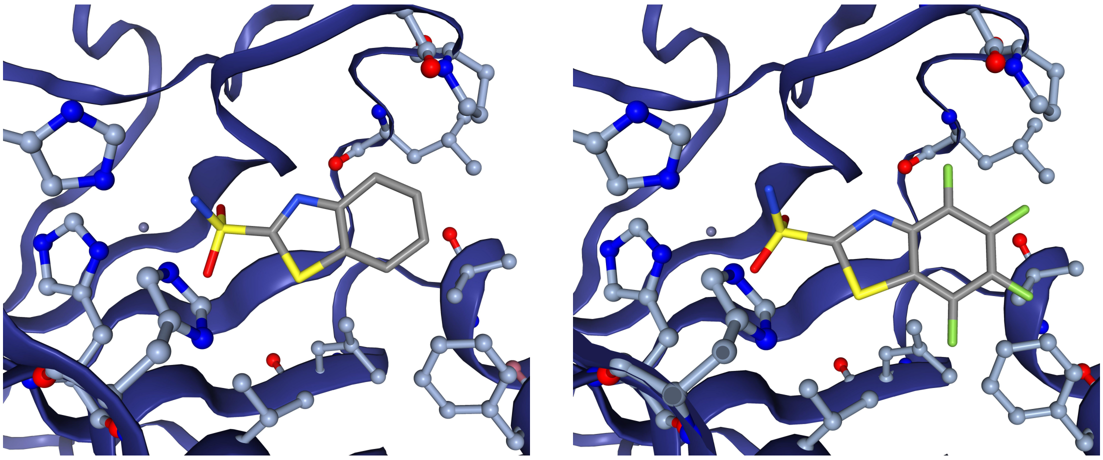
Enthalpy/Entropy Compensation (H/SC)

Enthalpy and entropy compensate each other, the stronger the binding the more negative the enthalpy and the lower the entropy. While in weaker binding, the entropy increases due to the system's disorder and thus the enthalpy decreases.

Often, the optimization of binding affinity was driven by increasing the Gibbs free energy ΔG. Due to enthalpy (ΔH)/entropy (ΔS) compensation (H/SC), more emphasis has been put on their separate contributions to the overall affinity.[77] For example the enthalpic optimization of an HIV protease inhibitor by introducing a strong hydrogen bond let to no overall gain in affinity due to an entropic loss attributed to conformational strain and solvation effects.[78]

Breiten *et al.* used human carbonic anhydrase (HCA) protein-ligand complexes with indistinguishable binding affinities for analyzing H/SC effects (Figure 1.4 and Table 1.1).[79] Herein, ITC was used to measure the changes in enthalpy and entropy, while WaterMap was applied to analyze the water molecules in and surrounding the active sites. They concluded, that apart from the ligand and protein, water molecules close to the interface have an impact on the enthalpy and entropy of binding.

However, the true impact of H/SC is controversial.[80,81] Uncertainties in experimental measurements of enthalpy and entropy exist. Additionally, other explanations apart from H/SC can often be found, such as solvation as an ubiquitous explanation or protein flexibility. Chodera[81] also showed, using an idealized protein-ligand binding site, that even though the conformational freedom of the ligand is restricted the tighter the ligand binds, and some H/SC is present, it is not linear as would be expected. Thus, they conclude that other aspects beside the H/SC must play a role.



(a) HCA with H_4BTA (PDBid 3s73[82]).

(b) HCA with F_4BTA (PDBid 4kap[83]).

Figure 1.4: Example of H/SC in HCA with two arylsulfonamide ligands; Figures were generated using the ProteinsPlus Server.[84]

Table 1.1: ITC measures for HCA (pKa corrected at 298.15K) taken from Breiten *et al.*.[79]

Ligand	ΔG°	ΔH°	-TΔS°
H_4BTA	-13.5 ± 0.3	-18.9 ± 0.5	5.5 ± 0.7
F_4BTA	-13.0 ± 0.2	-16.3 ± 0.6	3.4 ± 0.5

Hydrophobic Effect

The energetic gain due to the reorganization of water molecules upon protein-ligand binding is termed hydrophobic effect. Water molecules at hydrophobic sites of both protein and ligand are replaced, thus leading to an energy gain.[85] Upon release of water molecules close to hydrophobic areas of protein or ligand, conformational freedom would be gained. Thus, this energetic gain was mostly attributed to entropy changes. However, it has been shown that enthalpy changes can also contribute to the hydrophobic effect.[82,86–88] Experimentally, the energetic contributions of the hydrophobic effect range from -67 J mol^{-1} $Å^{-2}$ (with no temperature indicated),[89] to -125 – -138 J mol^{-1} $Å^{-2}$ and -119 – -149 J mol^{-1} $Å^{-2}$ (at room temperature).[90,91]

The question about the main driving force of protein-ligand association is still topic of discussion, as exemplified by studies from Setny[92,93] and re-interpretation of his studies by Graziano.[94] Setny constructed a model protein-ligand binding site (Figure 1.5).[92] Using MD simulations he concluded that the driving force of protein-ligand association was enthalpy-driven due to water reorganization. Then, Graziano derived a different interpretation of the same set-up of the protein-ligand binding site.[94] He concluded that the gain in enthalpy due to water-water hydrogen bond formation is compensated by the simultaneous loss in entropy. Thus, in his opinion, enthalpy cannot be the driving force of protein-ligand association. Again, Setny commented on the conclusions from Graziano, stating that his interpretation neglects dehydration effects of the binding site, which contribute to the overall thermodynamics.[93]

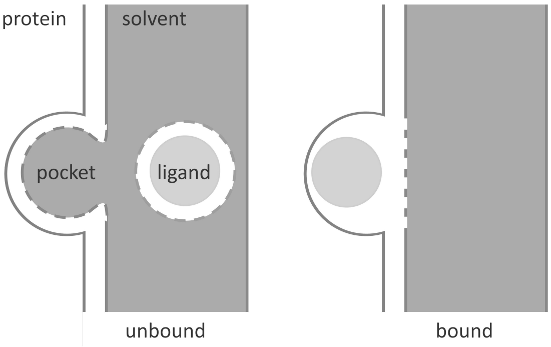


Figure 1.5: Model protein-ligand binding pocket used by Setny[92,93] and Graziano;[94] The dashed line indicates the solvent accessible surface area relevant for the shown binding event.

Kinetics

Water molecules not only play a significant role in the thermodynamics of e.g. protein-ligand association, but they also influence their kinetics, i.e. the residence time of a ligand in a binding site. Classically, IC_{50}, the necessary concentration to inhibit half the enzyme activity, and k_D, the dissociation constant, are frequent measurements for binding *in vitro*. However, especially *in vivo* the time-scale of protein-ligand interactions influences the magnitude as well as the duration of a response.[95–98] Therefore, the individual k_{on} and k_{off} rates are of interest. The importance of kinetics for drug development can also be seen in recent efforts by a public-private consortium approaching open questions such as transition from *in vitro* to *in vivo* or standardized data formats.[98]

Schmidtke *et al.* analyzed heat shock protein (Hsp) 90 inhibitor complexes for which detailed thermodynamic and kinetic data were available.[99] They concluded, that buried hydrogen bonds, i.e. polar atoms that are almost unaccessible, are exchanged at a lower rate. Thus, a hydrogen bond that is highly inaccessible for water molecules can be a rate-limiting step in protein-ligand binding. Bortolato *et al.* evaluated the role of water molecules for k_{off} rates.[100] Using a combination of different computational methods, Adenosine A_{2A} receptor antagonists were examined. Herein, unfavorable water molecules, also called 'unhappy' water molecules, included in protein-ligand binding sites influence off-rates. The number of unhappy water molecules as well as their actual positions in the binding site are of relevance. MD simulations were also applied to calculate kinetic parameters computationally.[101,102] For more information on the use of MD simulations for generation of kinetic data please refer to Deganutti *et al.*.[103]

1.2 Computational Methods for Water Molecule Predictions

Over the last years, many different methods for the characterization and prediction of water molecules became available. On the one hand, they can be separated according to the aim of the method:

A1 Classification of conserved vs. non-conserved waters,

A2 Estimation of the relevance of a water molecule,

A3 Prediction of water locations alone or in combination with their energetic contributions.

On the other hand, the different methods can be classified based on their underlying method:

M1 Empirical/Knowledge-based methods,

M2 Statistical/Molecular mechanics methods,

M3 Molecular dynamic simulation methods, and

M4 Monte-Carlo simulation methods.

In most cases, empirical/knowledge-based methods (*M1*) are applied to different aims, especially to points *A1* and *A2*, while the prediction of water locations and their thermodynamic profiles (*A3*) are often employed by methods *M2*, *M3*, and *M4*.

1.2.1 Underlying Technologies

In the following paragraphs, methods and theories applied by the water prediction tools are briefly introduced.

Free Energy Perturbation (FEP)[104–107] calculates the relative free energy between two systems using the Zwanzig equation:[104]

$$\Delta G_{A\rightarrow B} = -kT\ln\left\langle \exp\left(-\frac{\Delta E_{AB}}{kT}\right)\right\rangle_A \tag{1.2}$$

Herein, system A is the reference state, while system B displays the perturbed state. T is the Temperature and k is the Boltzmann constant. The energy difference between the two states $\Delta G_{A\rightarrow B}$ is the average of the ensemble generated by system A. Energy convergence can only be achieved if system A and B have a large enough overlap and are relatively similar. The latter is ensured by applying a reaction coordinate λ, which resembles the reference state ($\lambda = 0$) and the perturbed state ($\lambda = 1$). Any value in between describes a non-physical hybrid of both states. In FEP applications, the reaction coordinate λ is split and multiple λ windows are sampled.

Thermodynamic Integration (TI)[108–110] generates a potential of mean force (PMF) using the gradient of free energy over the aforementioned reaction coordinate λ, denoted as $(\delta G/\delta\lambda)_\lambda$. This free energy surface along the coordinate is calculated at different points of λ. The full PMF is generated by numerical integration of the different λ points. An alternative approach, finite-difference thermodynamic integration (FDTI), uses an approximation $(\Delta G/\Delta\lambda)_\lambda$ to obtain finite differences. Thus, the differences between the reference states at λ and the perturbed state at $\lambda + \Delta\lambda$ can be calculated using the Zwanzig equation (1.2). An advantage compared to FEP is the improved overlap between the reference and perturbed state by using $\Delta\lambda$. Therefore, the Zwanzig equation converges better. In order to verify the approximation, the energy for the forward $\lambda \rightarrow \lambda + \Delta\lambda$ and backward $\lambda \rightarrow \lambda - \Delta\lambda$ windows can be calculated with their absolute values being equal.

Inhomogeneous Fluid Solvation Theory (IFST)[111,112] decomposes the solvation free energy into four subunits: the solute–solvent energy, the solvent reorganization energy, and two entropic terms – the solute–solvent entropy and solvent reorganization entropy.[70] Thus, the effect on the free energy due to a solute is calculated relatively to its bulk state. One advantage of IFST is the detailed differentiation of energy contributions. On the other hand and directly related to this aspect is the high amount of approximations that have to be made in combination with rather complicated calculations. A recent study by Huggins and Payne resulted in a slight overestimation of entropic contribution but an overall agreement between hydration free energy predictions.[113]

Cell Theory (CT)[114] describes a method to calculate free energies of water molecules from computer simulations. Each water molecules resembles a rigid body with hindered translations and hindered rotations. Using computer simulations, potential energy surfaces are approximated. Thus, equilibrium simulations of water allow the calculation of its free energy, enthalpy and entropy. The entropy term can further be separated into translational, rotational, and conformational contributions.

Poisson-Boltzmann (PB)[115–117] *and Generalized Born (GB)*[118–121] are implicit solvation models. The advantage compared to explicit solvation models is the lower computational cost. PB resembles the electrostatic environment of a solute in a solvent containing ions. GB represents an approximation of the PB equation and is therefore computationally less expensive. Herein, the solute is modeled as spheres with dielectric constants different from the surrounding solvent. The combination of PB or GB with a surface area (SA) term allows the energetic penalty when cavities in the solvent are generated. Both PB/SA and GB/SA are too inaccurate to resemble single water molecules. This is partly overcome by the implicit solvation model AGBNP2 by Gallicaccio *et al.*.[122] For more detailed information of implicit solvent models and their application please refer to the following publications.[123–125]

Apart from implicit solvation models, diverse explicit solvation models exist. Herein, Hess *et al.* have conducted a study analyzing thermodynamic properties predicted for 13 amino acid side chains using three different force fields (AMBER99,[126] GROMOS 53A6,[127] and OPLS-AA[128]) in combination with five different water models (SPC,[129] SPC/E,[130] TIP3P,[131] TIP4P,[131] and TIP4P-Ew[132]).[133] They concluded, that different force fields lead to small variations in the calculated thermodynamics, while the water model had a greater impact on the calculated energies. For more details on implicit and explicit water models and their (dis-)advantages please refer to Onufriev *et al.*.[134]

1.2.2 Software Solutions

Knowledge-based Methods

Diverse empirical methods were developed to evaluate crystallographically observed water molecules and classify them into conserved/bound water molecules and non-conserved/displaceable ones.[135–138] Consolv,[135] WaterScore,[137] and PyWATER[138] mainly exploit environmental factors, such as B factor and the number of hydrogen bonds, for this purpose. WatCH[136] on the other hand identifies conserved waters in related protein structures using a hierarchical clustering approach.

Other knowledge-based methods determine whether a crystallographically observed water molecule is structurally relevant. Proasis WaterRank[139] is a purely geometry based score, consisting of hydrogen bond distances and angles of the crystallographic water molecule to its surrounding protein partners. Herein, a maximum of two donors and two acceptors is allowed. The Relevance metric[140] is a combination of WaterRank and HINT score.[141] Thus, it displays a combination of a geometric score with the interaction strength of a water molecule to its surrounding hydrogen bond partners.

Apart from characterizing crystallographically observed water molecules, methods have been developed that predict water positions and their corresponding energetic scores. Knowledge-based methods are used to place water molecules, mostly focusing on structurally relevant ones,[142–147] but also to evaluate the energetic contribution of the predicted water positions.[148–150]

AQUARIUS[142,143] and WATGEN[144] rely on observable water positions in crystal structures around amino acids. Those positions are transfered to the protein structure of interest and are used for water placement. AcquaAlta[145] has a similar approach as the previous two methods, but exploits the information of small molecules contained in the Cambridge Structural Database (CSD).[151] Herein, interaction geometries for water molecules with different functional groups are analyzed. The ideal interaction directions of protein and ligand are determined. They are scored and rank-ordered by *ab initio* calculations. The ranking is thus used to place water molecules. WaterDock[146] uses AutoDock Vina[152,153] to dock water molecules and thus identify suitable positions. A probabilistic water molecule classifier is applied to predict the conservation of water sites and to identify what type of atom, i.e. polar or apolar, can be used for its displacement. Xiao and coworkers developed a tetrahedral water-cluster model, which makes use of amino acid triplets to derive feature triangles.[147] The water molecule is then placed at the top of a tetrahedron of which the amino acids form the bottom triangle.

The water placement procedure in Fold-X, a force field method,[149] is based on the AQUARIUS approach, using known water positions to generate new water positions in the protein-ligand interface of interest. Those positions are then incorporated in the Fold-X force field and finally lead to water positions that are included in the calculation of free energy changes of side chain mutations. The HINT (hydropathic interactions) toolkit[148] is a grid-based approach. Herein, available grid points in a protein-ligand interface are scored with HINT.[141] The most favorable sites are then used to place a water molecule. The procedure is repeated with the already placed water molecules to account for bridging water interactions. DOWSER++[150] is the combination of DOWSER,[154] AutoDock Vina[153] and WaterDock.[146] Herein, internal cavities are detected and filled with water molecules.

Placing and/or scoring of water molecules is achieved using different methods: Statistical and molecular mechanics, molecular dynamic (MD) simulations, or Monte-Carlo simulations.

Statistical and Molecular Mechanics Methods

GRID[155–157] uses different probes to generate molecular interaction fields. Among these probes, one mimics a water molecule. The entropic contributions are estimated indirectly using a so-called 'dry' probe, which is evaluating the hydrophobic effect. Similar to GRID, MCSS (multiple copy simultaneous search)[158–160] uses different probe molecules, i.e. a water probe, to generate maps with their preferred locations. However, in contrast to GRID, MCSS does not use a grid but randomly distributes the probe molecules. Those positions are then minimized to retrieve energetically preferred positions. WaterFLAP[161] is based on GRID. Potential water positions identified with GRID are optimized locally and incorporated into the target structure. This procedure is repeated iteratively in order to not generate multiple water layers. An additional probe 'ENTR' is used to reflect the entropy of the water molecules. Using this probe, bulk-like water molecules are identified. Furthermore, the GRID 'CRY' probe is used to correct for hydrophilic regions. 3D-RISM (3D reference interaction site model)[162,163] generated 3D solvent site profiles. Using integral-equation theory of liquids, equilibrium solvent distributions can be obtained rapidly without sampling. Favorable hydration sited with localized entropies, enthalpies and solvation free energies are revealed by the distributions. wPMF (water potential of mean forces)[164] generates atom pair potentials based on water and 40 protein atom types. It is trained on nearly 4 000 protein structures to predict hydration sites and assign wPMF scores. SZMAP (solvent Zap mapping)[165] is the only tool using a grid-based semi continuum approach. One explicit water molecule (a so-called water probe) is sampled while the surrounding is simulated with a Poisson-Boltzmann continuum model (section 1.2.1). The thermodynamic properties are calculated using the water probe, a neutral probe, i.e. with charges removed, as well as a vacuum probe, with van der Waals terms removed. Setny *et al.* describe a semiheuristic solvation model based on body-centered cubic (BCC) lattice.[166–168] A water molecule is centered on a lattice point such that its neighbors are arranged in a tetrahedral arrangement. The hydrogens of the water molecule can occupy two of its eight direct neighbors, which results in a total of 12 possible water orientations. The energy of a water molecule at a lattice point is calculated using a position dependent effective Hamiltonian that includes solvent-solvent and solute-solvent terms. Iteratively, the lattice configuration – either occupied by a water or vacant – is determined using the effective Hamiltonian. This method is applied to predict buried water molecules in protein cavities.

MD Simulation Methods

WaterMap[169,170] uses a short MD simulation to sample the available space within a protein. The generated water positions are clustered and then IFST (section 1.2.1) is applied to calculate the thermodynamic profiles of the generated water positions. Similar to WaterMap, STOW (solvation thermodynamics of ordered waters)[171] uses an MD simulation in combination with IFST. Herein, mean energetic interactions of specific water positions are provided. Applying rotational and translational restrictions allows the calculation of an entropic penalty. SPAM ('Maps' spelled backwards)[172] applies a nanoscale MD simulation with explicit solvent. Using a site partition function, free energies are calculated. The site partition function is based on local distribution relative to the perturbation in bulk water. Water-water contacts are neglected in SPAM. WATsite[173,174] exploits clustered MD trajectories. The probability densities of translation and rotation of water molecules are used to estimate entropies. In general, no water-water contacts are considered. GCT (grid cell theory)[175–177] analyzes explicit solvent molecular simulations based on the cell theory method (section 1.2.1). Herein, interaction energies from water-solute to water-water are compared to retrieve thermodynamic properties. From a single simulation, insights of the enthalpies and entropies of hydration are generated. WATCLUST[178] predicts water sites that can subsequently be integrated into docking. An MD simulation is run to determine structural and thermodynamic properties. BiKi Hydra[179] analyses the persistence of water molecules using an MD based analysis. Steered MD simulation are run, followed by a spatial density analysis to measure the local water stability. Persistence of water molecules can be influenced by favorable interactions as well as steric hindrance. GIST (grid inhomogeneous fluid solvation theory)[180,181] is a discretization of IFST onto a 3D grid. The use of a 3D grid allows the calculation of water properties at every voxel of the grid. This way, explicit water locations are dispensable and water properties are provided as a function of position. MixMD (mixed molecular dynamics) has previously been used to identify active sites as well as allosteric sites in protein structures[182] and was recently applied to predict the displaceability of water molecules.[183] Six different probes, representing different chemical properties, can be tested for potentially displacing water molecule. The great advantage of this method is the knowledge about which probe is best able to displace a water molecule. This information can potentially be used for rational-driven chemical alterations of a ligand.

Monte-Carlo Simulations

RETI (replica exchange thermodynamic integration)[184,185] aims at calculating relative hydration free energies. Herein, a combination of FDTI and Hamiltonian replica-exchange method to enhance the sampling is used. In addition to the above described TI (section 1.2.1), multiple replica of the system are generated with different states being exchanged, when the Replica Exchange test is passed. Thus, an ensemble distribution of the reaction coordinate λ is generated. Double decoupling method[186,187] simulates two different states: Once, the decoupling of water from bulk, second, the decoupling of water from the receptor site. The latter is restrained to the active site of the protein to limit the sampling space to achieve convergence. Using the differences, free energies can be calculated. Double decoupling with RETI[188] is a combination of the two previously explained method. The thermodynamic properties are calculated using RETI for both decoupling processes. JAWS (just add waters)[189] also uses the double decoupling method to calculate energies. Herein, the conformation of the protein chain is sampled, while water molecules appear and disappear on a grid. Thus, hydration site occupancies can be calculated and incorporated into the interaction energy estimations. MCRS (Monte Carlo reference state)[190] simulates atom-atom contact densities by sampling the structural space with random probes. The aim is to predict hydration sites in proteins and construct the reference state of a system. GCMC (grand canonical Monte Carlo)[191–193] methods try to circumvent the drawbacks of MD simulations, i.e. the lack of successfully filling protein cavities isolated from the bulk. Using random moves – translation and rotation as well as deletion and insertion – physical barriers are avoided. The number of particles, i.e. water molecules, is fluctuating depending on a defined chemical potential. This chemical potential is usually defined from *a priori* knowledge about the occupancy of the cavity. Ross *et al.* have developed GCMC to define the chemical potential from the simulation itself.[194] Finally, the relation between free energy of water and chemical potential allows conclusions of the water affinity at specific locations. AquaBridge[195] is a Monte-Carlo based method to identify bridging water molecules in protein-ligand interfaces. Those positions can subsequently be integrated into docking.

1.2.3 Evaluation and Comparison Studies

Most of the aforementioned tools were validated on limited amount of data and only the distances between the placed and crystallographic water molecules were reported. The authors of DOWSER++[150] suggested statistical criteria for water placement evaluation. The criteria are based on the number of true positive – water molecules placed close to crystallographically observed ones – and false positive predictions – water molecules placed without close crystallographic ones. Thus, the sensitivity, also called recall, and false discovery rate of the method were calculated:

$$\text{Sensitivity} = \chi = \frac{\sum \text{True Positive}}{\text{N}_{X-ray}\ \text{H}_2\text{O}} \tag{1.3}$$

$$\text{False discovery rate} = \mu = \frac{\sum \text{False Positive}}{\text{N}_{placed}\ \text{H}_2\text{O}} \tag{1.4}$$

$$\text{Reliability} = \rho = (1-\mu)\cdot\chi \tag{1.5}$$

$N_{X-ray}H_2O$ is the total number of crystallographically observable water molecules and $N_{placed}H_2O$ the total number of placed water molecules. The false discovery rate is difficult to interpret. The number of observable crystallographic water molecules is limited. This is due to the fact, that the protein crystal structure is only an average of multiple protein structures, but also that waters close to the macromolecule are spatially more confined than others. Further, water molecules are not necessarily in the focus during structure elucidation or might be placed by automatic placement tools. Additionally, the crystal contacts, i.e. further protein units of the crystal, would also need to be considered in case the area of the protein where water molecules are placed is affected by them. Using a combination of sensitivity and false discovery rate, Morozenko *et al.* calculate a reliability score ρ. They applied DOWSER++ to a data set from Sleigh *et al.*[196] consisting of conserved water positions from 14 high- to medium-resolution structures of oligopeptide binding protein bound to a tripeptide

(Lys-X-Lys). Based on their distance criteria of 2.0 Å between a placed and a crystallographically observed water molecule, they correctly placed 85% of the water molecules (χ = 0.85) while the number of false positives was kept low (μ = 0.2). Thus, they achieved a reliability ρ of 0.69. According to their evaluation, WaterDock[146] achieved a higher sensitivity (χ = 0.95). However, WaterDock placed more water molecules than crystallographically observable (μ = 0.93), which lead to an overall reliability ρ of 0.07.

In 2014, as part of the critical assessment of predicted interactions (CAPRI), a blind prediction of protein-protein interfacial water positions was performed.[197] Twenty groups that previously submitted docking positions for another CAPRI study – docking predictions for the complex of DNase domain of colicin E2 and Im2 immunity protein – were invited to predict interfacial water molecules for that target. A total of 195 different models was submitted and evaluated based on 35 water-mediated contacts. Herein, two aspects concerning the sensitivity of the models were analyzed: (1) fraction of recalled water mediated contacts (f^{WMC}(nat)) and (2) fraction of recalled water molecules (f^{W}(nat)(r)).

$$\text{Recall(WMC)} = f^{WMC}(nat) = \frac{n_p^{WMC}}{n_t^{WMC}} \tag{1.6}$$

$$\text{Recall(W)} = f^{W}(nat)(r) = \frac{n_{p-matched}^{W}(r)}{n_t^{W}} \tag{1.7}$$

n_p^{WMC} is the number of correct predicted water mediated contacts, n_t^{WMC} the number of crystallographically observed contacts, $n_{p-matched}^{W}(r)$ the number of predicted water molecules in a certain distance r to a crystallographically observed water molecule, and n_t^{W} the total number of observed crystallographic waters. Within 1.0 Å distance, a maximum of 40% of the crystallographic water molecules was matched in combination with a maximum recall of water mediated contacts of 60%. Only 6% of the submitted models had $f^{WMC}(nat)$ above 0.5, which means that only half the water-mediated contacts were re-created correctly.

The sensitivity of the methods due to the absolute coordinates of waters or protein atoms was only evaluated for SZMAP[165] and WatSite.[173,174] The sensitivity of SZMAP results were analyzed by shifting water positions up to 0.5 Å 200 times. The score variability of the water ensemble was analyzed. Depending on the system, ΔΔG varied between 4.3 and 45.8 kcal mol^{-1}. WatSite results have been evaluated concerning the impact of the used MD parameters as well as the variation of binding site residue conformations. Results showed that a 4ns MD simulation is long enough to achieve reliable water sites. However, variation of over 0.5 Å in amino acid conformations let to inconsistent hydration sites and predicted energies.

Very few comparisons of different software solutions exist. Many tools have been applied to different problems, as recently reviewed by Bodnarchuk[198] or collected in a perspective about water prediction methods.[199] However, a thorough comparison, i.e. the application of different computational tools on the same target structure is hard to find. A study by Bodnarchuk compared three different Monte Carlo based methods – JAWS, double-decoupling with RETI and GCMC – applied to N9 neuraminidase.[200] The three methods predicted consistent water energies for a single and isolated water molecule. From the three tested methods, the results of JAWS can be more difficult to interpret, especially when water networks lead to overlapping density contours. Bortolato *et al.* have combined different methods – WaterMap, SZMAP, GRID, GCMC, and WaterFLAP – for the evaluation of ligand residence time for A_{2A} receptor (see section Kinetics 1.1.3).[100] They created the water network with SZMAP, optimized it with a customized version of WaterMap and then predicted the water energies with SZMAP, WaterMap, GRID, and GCMC. Finally, they analyzed the water networks to draw conclusions about ligand residence time. Overall, they mainly attributed the different residence time to trapped unfavorable water molecules in the protein ligand binding sites. Another combination of tools – GRID, WaterMap, and SZMAP[201] and GRID, WaterMap, and WaterFLAP[202] – has been applied to studying the druggability of GPCR binding sites. Herein, the hydration differences were studied and related to the overall druggability of the active site. Not only the absolute number of energetically unfavorably contributing water molecules influences the active site druggability, but also their location and arrangement within the binding site. Explicit water networks to derive conclusions about energetics and kinetics display a 'third dimension' in drug design strategies. Recently, a comparison of SZMAP, WaterFLAP, 3D-RISM, and WaterMap was published.[203] The four water prediction methods were used for predicting the energetic contribution of water molecules in three target proteins. The water energies were used to explain experimentally observed structure activity relationships (SARs) that could not be explained by protein-ligand interactions themselves. Overall, Buchner *et al.* concluded, that water energy prediction software can be useful to guide drug development studies. However, none of the methods outperformed any other method in this study. While WaterMap showed predictions in accordance with observed SAR, the predicted value of an unfavorably scored water molecule exceeded the experimentally observed energy upon displacement of the water molecule.

1.3 Relevance and Applications

Apart from the diverse computational tools specifically developed for the analysis and prediction of characteristics of water molecules, they play important roles in different other computational techniques such as virtual screening and docking. Moreover, multiple studies were conducted that analyze the consequences of water molecule displacement within a protein structure.

1.3.1 Water Molecules in Computer-Aided Drug Design

Diverse computational approached in computer-aided drug design have been augmented by the integration of water molecules. Most prominent are docking and virtual screening approaches. Herein, different strategies exist. Methods such as GOLD,[204] GLIDE,[205,206] and SLIDE[207] allow the treatment of explicit water molecules using a full atom representation. However, due to the computational cost, often only few water molecules can be handled explicitly. Other approaches such as FlexX[208,209] and FITTED[210–212] treat waters as spherical particles.

GOLD includes the scoring of water mediated interactions as well as water displacement during protein-ligand docking. GLIDE applies implicit solvation and the user can additionally select explicit water molecules. Further development, called GLIDE XP, includes water desolvation energy terms into the scoring function, i.e. a penalty for polar atoms that are insufficiently solvated as well as estimation of a penalty for water molecules with high amount of hydrophobic contacts. SLIDE is a docking algorithm that allows side chain flexibility during docking to account for induced complementarity of protein active site and ligand. Herein, conserved water molecules are considered based on predictions from Consolv (see 1.2.2). The conserved water molecules can translate during docking or are fully displaced. The displacement of conserved water molecules is integrated into the scoring function as a penalty term. Displacement of water molecules is penalized only if its due to apolar groups. Additionally, the penalty term is scaled using the confidence of conservation based on Consolv. FlexX allows user defined in- or exclusion of water molecules. Further development of the so-called 'particle concept'[209] allows explicit water placement. FITTED was developed for docking of flexible ligands into flexible proteins. The docking and virtual screening performance was evaluated for different water representations: (1) fixed integration of crystallographic water molecules, (2) displaceable integration of crystallographic water molecules, (3) placement of water particles, or (4) no water molecules at all.[212] They concluded that overall a flexible representation of the protein is advantageous in combination with displaceable crystallographic water molecules.

Multiple docking approaches are based on AutoDock.[152,153,213,214] One development explicitly for docking water molecules, WaterDock,[146] has already been mentioned in the previous section (see 1.2.2). Furthermore, water molecules have also been integrated into docking tools to enhance their original purpose – docking of ligands into protein binding sites. A combination of AutoDock and GIST incorporates thermodynamics of active-site water molecules into the scoring function for protein-ligand docking.[215] Herein, a GIST-based desolvation function was integrated into the AutoDock scoring function to account for the displacement of unfavorable water molecules. Their application to dock 52 ligands into coagulation factor Xa resulted in a higher scoring accuracy as well as better docking poses. They further applied their AutoDock-GIST function to a virtual screening of factor Xa from the directory of useful decoys-enhanced (DUD-E)[216] resulting in a higher enrichment and better area under the curve (AUC) values. Uehara *et al.* further assumed, that displacement energy might already be implicitly included in scoring functions due to their training on experimental binding affinity. Thus, they suggested that more effort should be put on an explicit water term to account for their displacement energies. For the analyzed water molecules in factor Xa they concluded that most of them were energetically unfavorable based on their enthalpic contributions rather than their entropic contributions. Another approach based on the AutoDock force field[217] was developed by Forli *et al.*[218,219] Herein, water molecules, represented as spherical particles with combined donor and acceptor functionality are attached to ligands prior to docking. During the docking process, their positions are iteratively evaluated. The force field has been adapted to account for a spherical water model to calculate and include enthalpic and entropic water contributions. The inclusion of water molecules into the docking process led to an accuracy increase of 10% and 11.7% for training and test set, respectively. Not only for protein complexes but also for RNA the inclusion of displaceable water molecules led to an increase in docking accuracy as shown by Moitessier *et al.*.[220]

Apart from the above mentioned combination of AutoDock and GIST, DOCK3.7 was combined with GIST to enhance the docking performance by including water-displacement energies.[221] Retrospectively, 25 targets of the DUD-E were used to calibrate the weights of DOCK3.7 and GIST terms. Then, a prospective docking run was performed for a model cavity in cytochrome c peroxidase. Among the top 1000 scored poses, more than 50% overlap between GIST vs non-GIST docking. However, the combination of GIST and DOCK3.7 showed a correct ability to prioritize molecules.

A study by Huang *et al.* used a dynamic on/off-switching of water molecules, which led to an increase in docking accuracy.[222] Another recently developed docking method and scoring function, WScore, is based on WaterMap calculations.[223] Herein, water locations and thermodynamics are derived from WaterMap calculations. An ensemble docking approach is applied to account for protein flexibility. Water molecules are treated flexible during docking, with the possibility of being displaced into bulk. Desolvation effects are assessed based on solvation of polar and charged groups of protein and ligand.

As shown by multiple studies,[212,224–227] the influence of water molecules on the docking accuracy is greatly influenced by the protein system. Other computational approaches where water molecules can be considered are 3D-QSAR models[228–231] as well as pharmacophore models.[232–235] More information about chemoinformatic strategies can be found in D1.

Many of the aforementioned techniques rely on the availability of apo- or holo-structures of the protein of interest. However, especially apo-structures are not necessarily available. Additionally, the implicit assumption that the water molecule positions do not significantly change between apo- and ligand-bound structure is not always true. The same applies for holo-structures. The binding mode of the crystallized ligand might be different from the ones used in the docking procedure. Therefore, water placement tools display an opportunity to place water molecules even if the binding mode changes or no apo-structures are available.

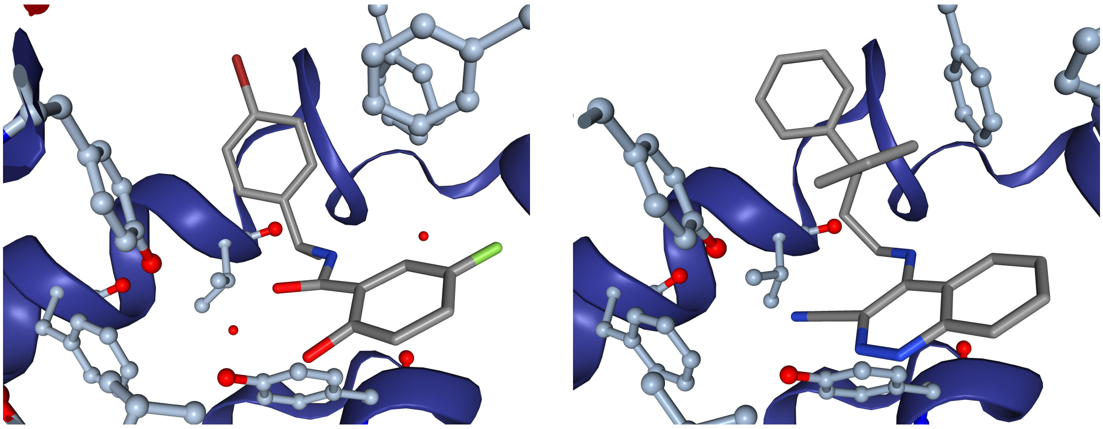
1.3.2 Water Molecule Displacement Studies

Diverse studies, computationally as well as experimentally, have been conducted to analyze the displaceability of water molecules.

Experimental Studies

In the following section, experimental studies on displaceability and relevance of water molecules will briefly be discussed. Generally, cyano groups,[20,21,236] halogens,[237] mimicry of the water molecule's interactions,[238] or methyl groups ('magic methyls'),[239] are exploited for water displacement strategies. For more details on the exploration of water molecules in drug discovery, please refer to a recent perspective[240] on the roles of water molecules.

Scytalone dehydratase (SD) is an enzymatic determinant for fungal disease targeting rice plants. A developed inhibitor[241] showed an extended hydrogen bond network with side chains and two water molecules (Figure 1.6a). Thus, another inhibitor series was developed aiming at displacing one of the water molecules to achieve an energy gain.[20] Herein, the water was displaced using a cyano group (Figure 1.6b), resulting in three proposed aspects concerning the thermodynamics: (1) entropy gain by water displacement into bulk, (2) less entropy loss due to conformational limitations, i.e. the ortho-CN substituent limits the conformational flexibility of the ligand, and (3) enthalpic gain due to interactions between inhibitor and protein. Inhibition constants due to different ligand alterations (Table 1.2) show that displacing the water molecule leads to the largest affinity increase. While a carbon atom is most unfavorable, due to no interaction with the water, a nitrogen atom is more favorable, probably due to a potential hydrogen bond interaction to the water molecule, while the cyano group, displacing the water molecule, shows the best inhibition.

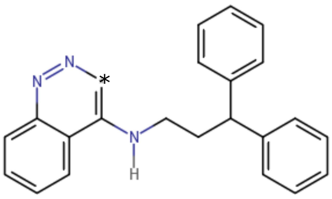


(a) SD with a salicylamide inhibitor (PDBid 1std[241]).

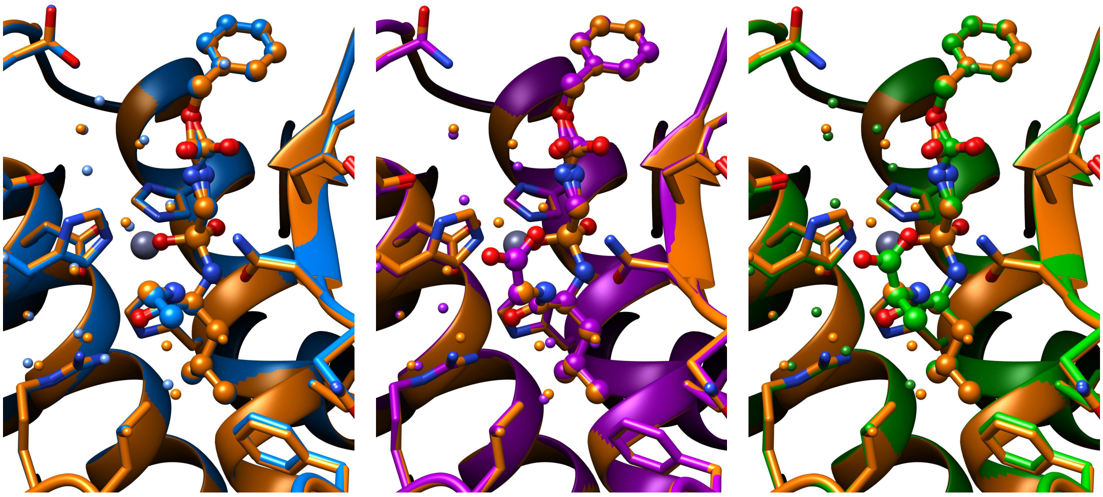
(b) SD with cyanocinnoline inhibitor (PDBid 3std[20]).

Figure 1.6: Example of water displacement in SD; Figures were generated using the Proteins*Plus* Server.[84]

Table 1.2: Inhibition constants of different SD inhibitors taken from Chen *et al.*;[20] * represents different ligand alterations listed in the table; ligand ids are taken from Chen *et al.*.[20]

	Ligand ID	*	K_i (nM)
	6d	C	140 ± 9
	5d	N	0.22 ± 0.02
	7d	CC#N	0.0077 ± 0.001

Non-additive effects of ligand alterations in thermolysin (TLN) were attributed to disruptions of the water network.[242] Herein, sequential addition of substituents, methyl and carboxylate, let to a moderate affinity gain for the first addition independent of which substituent was added first (Figure 1.7 and Table 1.3). However, the addition of the second substituent, again independent of which one, led to a larger affinity gain. The analysis of the water network surrounding the ligands, showed a disruption of the water network upon addition of the first substituent and a re-organization after the second addition.



(a) Superposition of **1** and **2**. (b) Superposition of **1** and **3**. (c) Superposition of **1** and **4**.

Figure 1.7: Example of water network rearrangements in TLN with different inhibitors;[242] Initial inhibitor **1** (orange, PDBid 3t73); Inhibitor **2** with methyl substituent (blue, PDBid 3t8f); Inhibitor **3** with carboxylate substituent (purple, PDBid 3t8g); Inhibitor **2** with both substituents (green, PDBid 3t74); Ligand IDs are taken from Biela *et al.*.[242] (Molecular graphics were created using UCSF Chimera.[65])

Table 1.3: ITC measures for TLN taken from Biela *et al.*.[242]

Ligand			ΔΔG°	ΔΔH°	-TΔΔS°
1	→	**2**	-2.2	+2.5	-4.7
1	→	**3**	-1.0	-3.5	+2.5
1	→	**4**	-6.7	-16.8	+10.2

DNA methyltransferase (DNMT) contains a highly integrated water molecule, which mediates the recognition of histone H3 with a trimethylated lysine (H3K36me3, Figure 1.8).[243] The recognition of DNMT and H3K36me3 initiates the methylation of DNA. Herein, a water molecule mediates between the serine side chain (Ser270) of DNMT and the trimethylated lysine backbone oxygen (3mLys36) of H3. A mutation of the DNMT serine to proline (Ser270Pro) leads to an affinity loss, thus a decrease in DNA methylation. The lack of methylated DNA leads to a disease called ICF syndrom (immunodeficiency, cantromeric instability and facial abnormalities). Rondelet *et al.* hypothesized that upon mutation of Ser270Pro the water mediated interaction is no longer possible, which leads to an affinity decrease. Another structural aspect that has been mentioned by Ge *et al.* is the tryptophane (Trp263) interacting with Ser270.[244] Upon mutation, Trp263 is no longer able to form a hydrogen bond with Ser270Pro and thus might need to adapt its conformation. Due to Trp263 being part of the aromatic cage necessary for the recognition of the trimethylated lysine, this might lead to an additional steric hindrance in binding of H3K36me3 by DNMT.

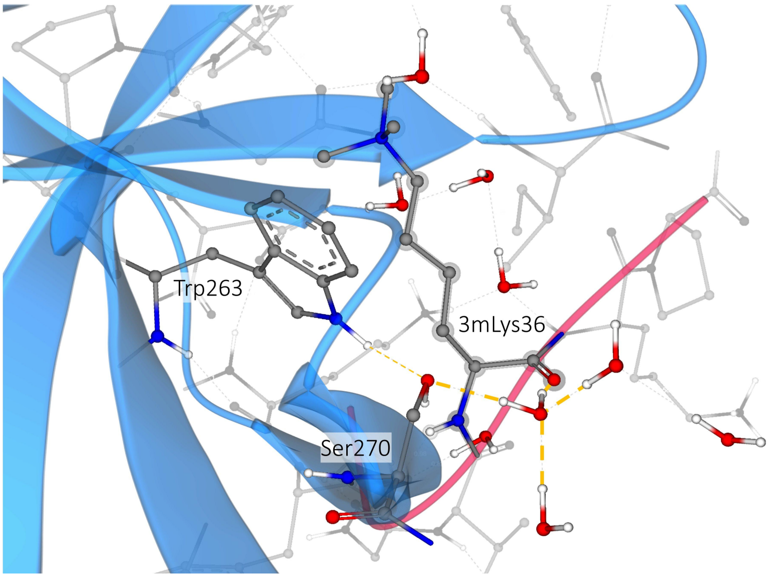


Figure 1.8: DNA methyltransferase (DNMT, light blue) with histone H3 recognition site (H3K36me3, red, PDBid 5ciu).[243] Water-mediated hydrogen bonds between Ser270 of DNMT and backbone oxygen of trimethylated Lys36 of histone H3 are indicated with yellow dashed lines.

Salie *et al.* examined the unusual long half-life of a HIV reverse transcriptase inhibitor.[245] They reasoned that this might be due to ordered water networks in the active site. The identification of four structurally relevant water molecules in the active site of β-amylase led to the conclusion that those waters keep eight side chains conformationally restricted.[246] These strategic water molecules can be displaced by ligand oxygen atoms. However, the number of displaced water molecules is determined by the substrate type, the remaining water molecules guide the correct orientation and hydrogen bond network in the active site necessary for the catalytic functionality of the enzyme.

Some water molecules in protein structures have proven to be difficult to displace or only without the expected affinity gain. Levinson *et al.* analyzed the significance of structural water molecules in Src kinases.[247] A pair of structured water molecules in the kinase active site is conserved among 40 different kinases with 164 ligands. They conclude that water mediated interaction can be critically exploited for specificity during drug development and should thus gain more attention throughout this process. Similar observations were made for HIV protease, where the exploitation of water-mediated interactions can be used for potent inhibitor design.[248]

Computational Studies

Different protein systems have been subject to computational studies about the displaceability of water molecules. Especially WaterMap has been applied to various protein targets. A summary of those applications can be found in a recent review.[249]

Factor Xa, a protein of relevance in blood coagulation, has been studied by WaterMap[170] as well as GIST[181] analysis. The WaterMap study concluded that enthalpic contributions resulted from displacing hydrophobically enclosed water molecules, while entropic contributions were due to displacement of hydrogen bonded water molecules. Additionally, they suggested to score water molecules not only based on the current complex, but to consider its contribution upon displacement. Energetically, the GIST study reached similar results, namely that energy may outweigh entropy in cases of strong hydrophobic binding, which was previously shown experimentally[86,88,169,250] as well as computationally.[92] Similar conclusions were drawn from an IFST study of 103 ligands in the ATP binding site of heat shock protein 90 (Hsp90), a molecular chaperone that aids other proteins folding correctly.[251] Haider *et al.* observed a better correlation between experimental binding affinity and predicted interaction energy than with predicted free energy. This is explained by less entropic contributions to the overall binding affinity.

An IFST calculation was done for a highly conserved water molecule in HIV protease, which cleaves so-called polyproteins to generate functional mature proteins.[252] Herein, to displace the water molecule, it was mimicked by an urea group. Compared to the carbonyl group of urea the water molecule is more favorable. Although the interactions of the carbonyl group are calculated to be slightly more favorable (-16.9 kcal/mol) than the free energy of the water molecule (ΔG_{solv}= -15.2kcal/mol), the desolvation of the carbonyl group (+5 kcal/mol) leads to an overall unfavorable contribution.

A TI calculation was applied to analyze the SH3 domain of Abl tyrosine kinase, a mediator of protein-protein interactions and therefore of relevance in cellular signaling.[253] Different functional groups were analyzed concerning the displacement of a tightly bound water molecule. Interestingly, hydroxyl, formamide, and an ethyl groups led to a favorable energy gain, while methyl and amine groups resulted in an unfavorable energy.

WaterMap analysis of A_{2A} receptor, a regulator of myocardial oxygen consumption and coronary blood flow, allowed the comprehension of affinity changes upon ligand alterations.[254] Herein, small apolar substituents led to a decrease in affinity, while larger hydrophobic substituents led to an increase. The water molecule analysis concluded that small substituents competed with favorable water molecules, thus leading to no affinity gain while larger substituents reached into an area with unfavorable water molecules. The displacement of those sterically unstable water molecules led to an affinity gain. Overall, these results show the importance of water molecule analysis, because in this case the differences in affinity would not have been explainable by effects such as ligand-receptor interactions or steric effects. Another study of the A_{2A} receptor identified an 'unhappy' water molecule as the reason for an unexpectedly large affinity gain (33-fold) upon addition of a methyl substituent – a so-called 'magic methyl'.[239]

A Concanavalin A study based on NMR experiments and MD simulations showed that a distorted water molecule contributed favorably.[255] Concanavalin A is a lectin for storage and defense mechanisms in plants. NMR studies showed that a hydroxyethyl moiety was not able to displace a conserved water, but led to a distortion. The authors suggested that the entropic gain (max 2.0 kcal mol^{-1}) would not outweigh the enthalpic loss upon displacement. This study contradicted previous conclusions from ITC calculations,[256] which were based on the displacement of the conserved water molecule.

As concluded previously by Garcia-Sosa[257] and shown by the above mentioned examples, it is difficult to predict the energetic outcome upon water molecule displacement. Diverse aspects have to be considered – water network effects, hydrophobic effect, enthalpic as well as entropic contributions among others. At the same time, these aspects show the relevance and importance of water molecules in drug design strategies.

1.4 Motivation and Thesis Content

Even though many aspects concerning the relevance of water molecules are already known, it is still an ongoing field of research. In order to understand binding aspects – protein-ligand as well as inter and intra protein – in more detail, it is necessary to understand the role of water molecules in greater detail. As shown in the previous sections, water is an integral part of the protein structure and has to be considered correctly.

Diverse areas of computer-aided drug design (CADD) could profit from a better understanding and integration of water molecules into software solutions, from docking, to scoring, to analyzing different water networks as a result of protein flexibility (D1). Frequently applied software solutions for placing and scoring water molecules are time-consuming. A dynamic application to large amounts of protein-ligand complexes is thus limited. At the same time, existing water prediction tools are hardly evaluated thoroughly. Further demands on water prediction tools are usability aspects such as easy-to-use and easy-to-interpret. If too many (too) close water molecules are predicted, the result is hard to interpret. In addition to this, the predictions should be consistent, i.e. if the crystallographically observable water network is not perturbed by ligand changes and the amino acids around the water molecules are not changed either, the predicted water locations as well as thermodynamics should be consistent. Furthermore, the method should – ideally – not depend on the protein system, but should be applicable universally.

Since it is hardly possible to retrieve binding affinity data for single water molecules experimentally, other means of validation needed to be considered. Protein structure models are an interpretation of the experimentally collected data, the diffraction patterns. The electron density data is exploited, since it is the only experimental data abundantly available for water molecules. The developed metric, called EDIA – **E**lectron **D**ensity of **I**ndividual **A**toms – allowed an automated comparison of the structural model with its underlying electron density (D2 and D3). This way, the foundation for validating the following water placement procedure was built.

In order to place water molecules, freely available space within protein structures needs to be identified correctly. Since water molecules participate in hydrogen bonds, a large-scale analysis was performed (D4, D5). Hydrogen bond geometries were derived from this study, which were used as a starting point for placing water molecules (P1).

In a final step, the predicted water positions were integrated into and scored with HYDE. Two aspects were considered concerning the estimation of the energetic contribution of water molecules: The energy value of the water molecule itself and in its surrounding. This gave valuable information about the 'happiness' of the water molecule. If a water molecule has a favorable energy contribution, it is unlikely to be displaced by a ligand. On the other hand, if the energetic contribution is unfavorable, this water molecule displays a possible means of increasing the protein-ligand binding affinity. Therefore, a suitable ligand is needed to displace the water molecule.

Finally, a comparison of state-of-the-art water programs (P2) as well as further application scenarios (D6, D7 and D8) were conducted.

2

EDIA – Structure Validation Using Electron Density

Water molecules display a great challenge when it comes to the availability of experimental data. As explained in Chapter 1.1.1 X-ray structures constitute a great means of resources and are readily accessible via the PDB.[23] However, just because water molecules are modeled in the X-ray structure does not necessarily mean that they are observed experimentally.[258] The X-ray structure is already an interpretation of the X-ray diffraction pattern collected through exposure of protein crystals to X-ray beams (Figure 2.1).

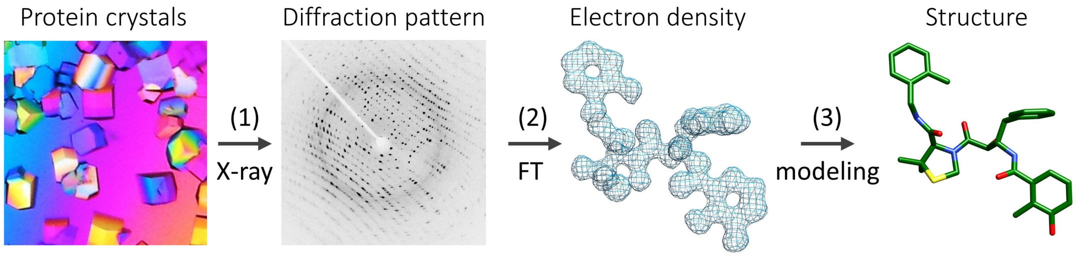


Figure 2.1: Structure elucidation process: (1) protein crystals[259] are placed in an X-ray beam and diffraction patterns[260] are collected; (2) generation of electron density model from diffraction pattern using Fourier transformation (FT); (3) molecular modeling of the atomistic structural model into the electron density.

The electron density, which is nowadays available for a large amount of protein structures, was exploited for the validation of water molecules. Using the electron density a metric for automatic evaluation of atoms with their underlying electron density was developed – EDIA (Electron Density of Individual Atoms). First, EDIA was developed solely for water molecules, i.e. a single oxygen atom. A more detailed description of the EDIA calculation can be found in D2. Then, EDIA was advanced to handle multiple atoms, such as ligands, amino acids, or whole proteins, called EDIA$_m$. For more details on EDIA$_m$ see D3. In this chapter the basic methodologies of EDIA and EDIA$_m$ are explained.

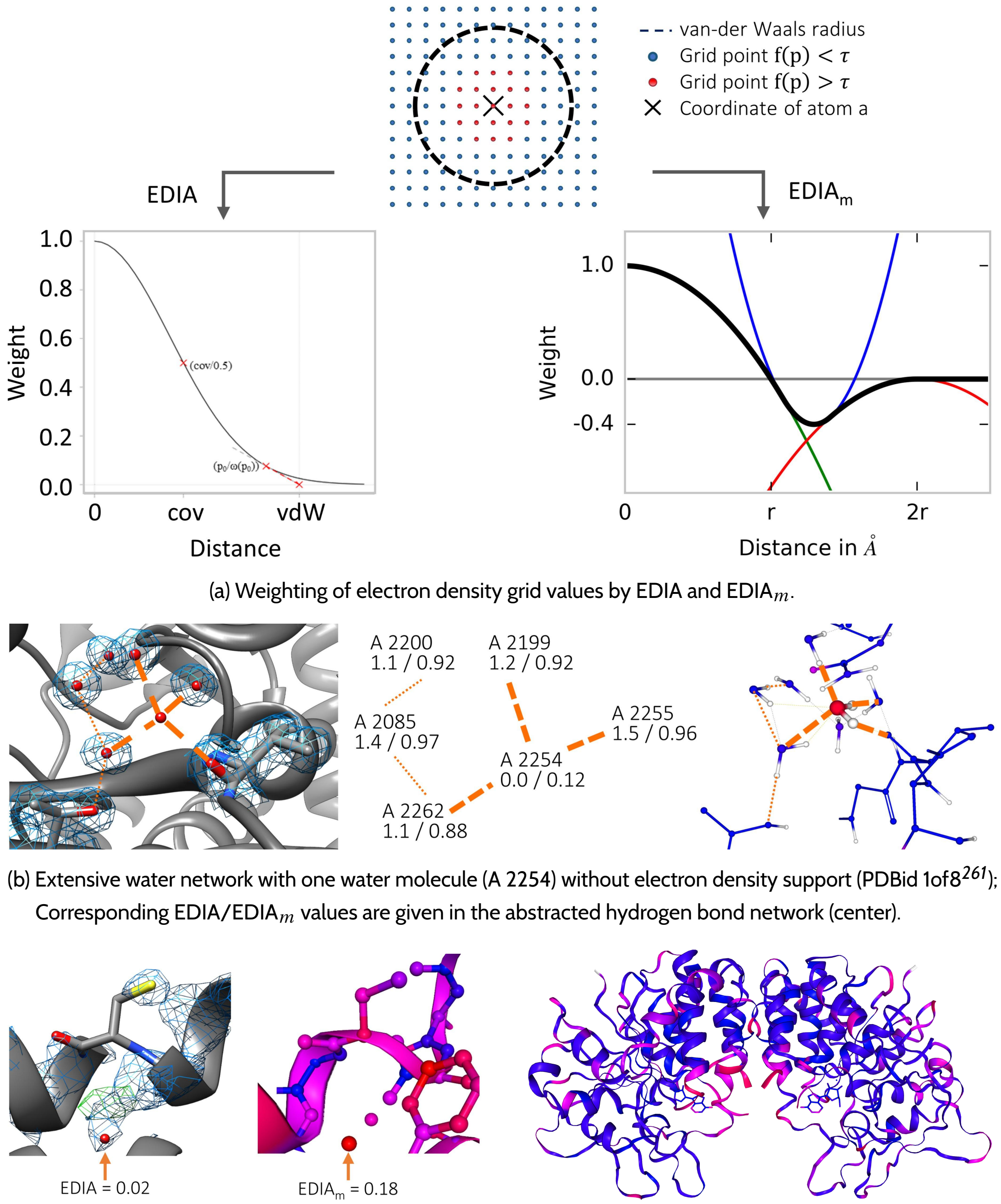
2.1 General Overview

The EDIA was developed as a means of validation for the water placement and water scoring procedures developed in the course of this dissertation. Herein, an objective and – more importantly – automatic measurement for the quality of water molecules was needed.

As already mentioned in Chapter 1.1.1, different measures exist for the evaluation of protein structures. However, existing measures such as B-factor, RSR, or RSCC were either not sufficient or too elaborate for our application. The B factor depends on the structure refinement and can thus be influenced. Additionally, it does not contain information about the underlying electron density. Other measures, such as RSR or RSCC, include the back calculation of an electron density grid from the modeled structure to compare it to the experimentally observed electron density. However, they do not consider clashes or unaccounted electron density. From a modeling perspective, the fit of the given structural model into the electron density is of interest. Thus, EDIA was developed which quantifies the fit of the given model into the available electron density.

The EDIA calculation was modeled by a Gaussian-like function, in which the electron density grid values closer to the center of the respective atom were assigned a greater weight, than density grid values at the outer radius (Figure 2.2a *EDIA*). Thus, water molecules could be classified based on the underlying electron density.

The EDIA was developed further to being able to handle multiple atoms, called EDIA$_m$. Different aspects have to be considered, when electron density coverage for multiple atoms is calculated. Especially neighboring effects such as covalent bonds, clashes, or unaccounted electron density have to be considered. These effects were integrated into the EDIA$_m$ calculation by putting great emphasis on the assignment of electron density ownership, i.e. the mapping of electron density to individual atoms. This way, the single electron density grid points (Figure 2.2) were assigned to one or – in case of covalent bonds – to multiple atoms. Based on the assigned ownership, electron density grid values are considered up to twice the radius of an atom. While electron density values are mandatory up to the full radius of the atom, electron density at further distances is unwanted and can be used as an indicator of problematic structural aspects (Figure 2.2a *EDIA$_m$*). Using the EDIA$_m$ whole protein structures, single amino acids, or individual atoms can be quantified based on their underlying electron density support.



(a) Weighting of electron density grid values by EDIA and EDIA$_m$.

(b) Extensive water network with one water molecule (A 2254) without electron density support (PDBid 1of8[261]); Corresponding EDIA/EDIA$_m$ values are given in the abstracted hydrogen bond network (center).

(c) Electron density of an incorrectly modeled methionine (Cys B 85) and water molecule (B 1497) with corresponding EDIA and EDIA$_m$ values (PDBid 1hp0[262]).

(d) Protein secondary structure colored by EDIA$_m$ (PDBid 1hp0[262]).

Figure 2.2: Calculation of EDIA and EDIA$_m$ and example structures; EDIA$_m$ coloring scheme: blue (well resolved) to red (not sufficiently supported by density).

2.2 Validation of Water Molecules

The initial application of the EDIA was to analyze the characteristics of water molecules. Therefore, a high-resolution PDB subset was compiled, with resolutions better than 1.5 Å. Even for high-resolution structures, we were able to show that almost 9% of all water molecules were insufficiently resolved. The majority of these unresolved water molecules was located at the protein surface, where water molecules are in general more flexible. Due to the atomic structure being an average of multiple proteins in the protein crystal, non-localized water molecules lead to a less clear signal. However, also water molecules close to protein-ligand or protein-protein interfaces were found to be unresolved, 3.8% and 3.3%, respectively. Overall, this shows the importance to distinguish water molecules and critically reviewing the data provided in X-ray structures.

With the development of EDIA we are now able to differentiate between water molecules that are well resolved by electron density and those that are not (Figure 2.2b). The compiled high-resolution data set provides a basis for the validation of further method development throughout the following chapters.

2.3 Validation of Multiple Atoms

The developed EDIA$_m$ allows the validation of X-ray structures as well as the evaluation of computational methods such as docking or geometric optimization prior to scoring. For our intended application, small molecules, amino acid side chains, or whole proteins are subject to a geometric optimization, called GeoHYDE. After optimization, the structure should still be in agreement with its underlying electron density. Thus, an automatic evaluation is necessary as a quality measurement – EDIA$_m$.

The developed EDIA$_m$ can be used for both, the validation of single atoms (Figure 2.2b) as well as the analysis of amino acid side chains (Figure 2.2c) or whole proteins (Figure 2.2d). The EDIA$_m$ also allows to automatically compile data sets with well resolved structures. Herein, analysis showed that out of 45,113 ligands from PDB structures 77% were well resolved (EDIA$_m$ ≥ 0.8). In a detailed comparison with other existing measures, such as B factor and RSCC, the advantages of EDIA$_m$ could be shown. Especially the automatic error detection, such as clashing atoms, too few or too much electron density, or shifted electron density, can aid a more detailed structural understanding.

Overall, the EDIA$_m$ is well suited for both requested tasks, structure validation as well as quality criterion for structure optimization purposes.

3

*NAOMI*nova – Analysis of Interaction Geometries

A critical step for placing water molecules is the correct identification of areas within protein complexes where water molecules would fit. Since most water molecules form (multiple) hydrogen bonds in protein structures, the idea was to exploit unsatisfied hydrogen bond functions[a] from ligand or protein atoms as a starting point for water placement. Therefore, different interaction types and their corresponding interaction geometries were analyzed.

IsoStar,[263] is a commercially available tool for the analysis of atom distributions around a central functional group. Herein, predefined functional groups are available. However, IsoStar was too strict in its functional group definitions compared to the objectives of this analysis. SuperStar[264,265] is another tool for the analysis of 'hot spots' in protein-ligand interfaces. It is based on data derived from IsoStar and allows the identification of favorable areas within a protein-ligand interface for specific functional groups. Both, SuperStar and IsoStar, do not allow a sufficient geometric analysis of the atom distributions with the flexibility needed for the purpose of this analysis.

Thus, the tool *NAOMI*nova was developed. For more information on the methodical details see D4 and for its application to hydrogen bond interaction definition see D5. In this chapter, the underlying method of *NAOMI*nova is briefly explained and its applications are presented.

[a] A hydrogen bond function is a donor hydrogen or an acceptor lone pair.

3.1 General Overview

*NAOMI*nova is a tool for geometrically analyzing atom distributions around a substructure, e.g. a functional group of interest. Herein, great emphasis was put on the flexibility of the analysis opportunities. The user can define the data for which an analysis should be performed, either a subset from the PDB, in-house data, as well as MD trajectories. The substructures can be determined by either using a SMARTS expression or by visually selecting atoms from a small molecule. Finally, the analysis of the data can be filtered with diverse geometrical as well as chemical criteria. A general overview of the data preparation as well as data analysis process is given in Figure 3.1.

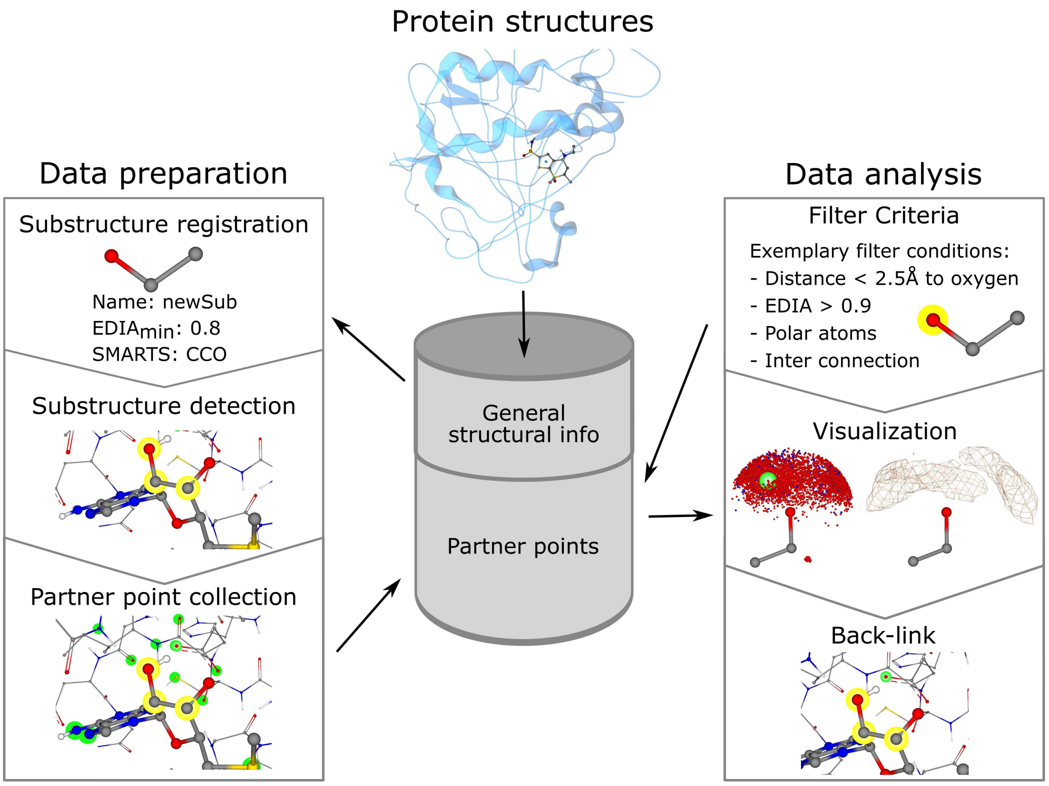


Figure 3.1: Overview of *NAOMI*nova data preparation (left) and data query (right) process. *Protein structures*: A user-defined selection of protein structures is used as input for the *NAOMI*nova database. *Data preparation*: Structural information is stored in the database and a user-defined substructure can be defined. This substructure is searched in the given structures and according to specifications of the user, i.e. EDIA$_m$ or SMARTS for a description of the immediate surrounding. Atoms within 4.5 Å, so-called partner points, are stored in the database for later querying. *Data analysis*: Based on diverse chemical as well as structural criteria, the database can be queried for stored partner points. The substructure in combination with the partner points can be analyzed visually for geometric characteristics as well as their underlying structures using the integrated back-link option.

Apart from the geometric analysis of data, *NAOMI*nova also provides the option of analyzing a protein-ligand binding site of interest. Herein, suitable filtered data can be super-imposed in the protein-ligand active site. Thus, preferred interaction directions or assemblies of surrounding atoms can be visualized and analyzed. For applications of binding site analysis please see D4.

3.2 Analysis of Hydrogen Bond Geometries

The development of *NAOMI*nova allowed the analysis of diverse interaction types in more detail. Additionally, *NAOMI*nova provided the necessary flexibility for further geometric analysis of hydrogen bond interaction geometries.

Two main purposes were subject of this analysis: (1) the definition of preferred interaction geometries and (2) the verification of the used interaction geometries in the NAOMI software library.[266–268]

Evaluating Hydrogen Bond Geometries

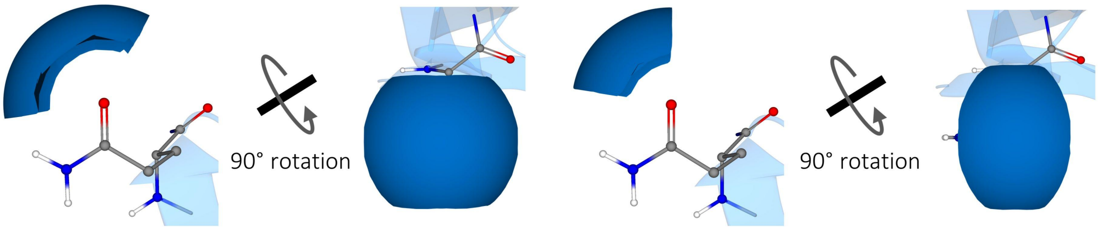
For our evaluation of hydrogen bond geometries, we focused on the evaluation of atom distributions around a functional group within ideal hydrogen bond distances (2.6 – 2.9 Å). Some publications mention a more linear geometry at shorter hydrogen bond distances[269,270] or in other words, wider distributions of partner atoms at larger distances. The main problem at larger distances, especially greater than 3.0 Å, is the distinction between 'true' hydrogen bonds and coincidental contacts. Thus, we kept the distance for our analysis between 2.6 Å and 2.9 Å to focus on relevant hydrogen bond geometries.

Definition of Hydrogen Bond Geometries

Diverse applications rely on hydrogen bond geometries: most existing scoring functions and also the throughout this dissertation developed water placement procedure. Since the defined geometries are used for scoring the quality of hydrogen bonds, it is assumed that the number of observations relates with their energy contributions.

With *NAOMI*nova 22 different functional groups were analyzed. Especially for sp^2 hybridized oxygen atoms, such as oxygen atoms of a carbonyl, amide, or ester, differing geometries were observed. Those differences between sp^2 oxygen atoms were not expected. Additionally, the geometries varied from those used within the NAOMI library (Figure 3.2). Due to a rectangular geometry, modeled as a spherical rectangle, the new geometry (Figure 3.2b) is rotated by 90° compared to the previous definition (Figure 3.2a).

For more information on the briefly described evaluation and definition of hydrogen bond geometries please refer to D5.



(a) 'Old' geometry definition.

(b) Experimentally derived geometry.

Figure 3.2: Spherical rectangle geometry of a primary amide oxygen; only one interaction direction is shown by the blue surface.

3.3 Water-Surrounding Analysis

Further, *NAOMI*nova was applied to the analysis of distances between water molecules and their surrounding atoms. This information was needed for the water placement procedure where the overlap of surrounding atoms with a water molecule was of interest.

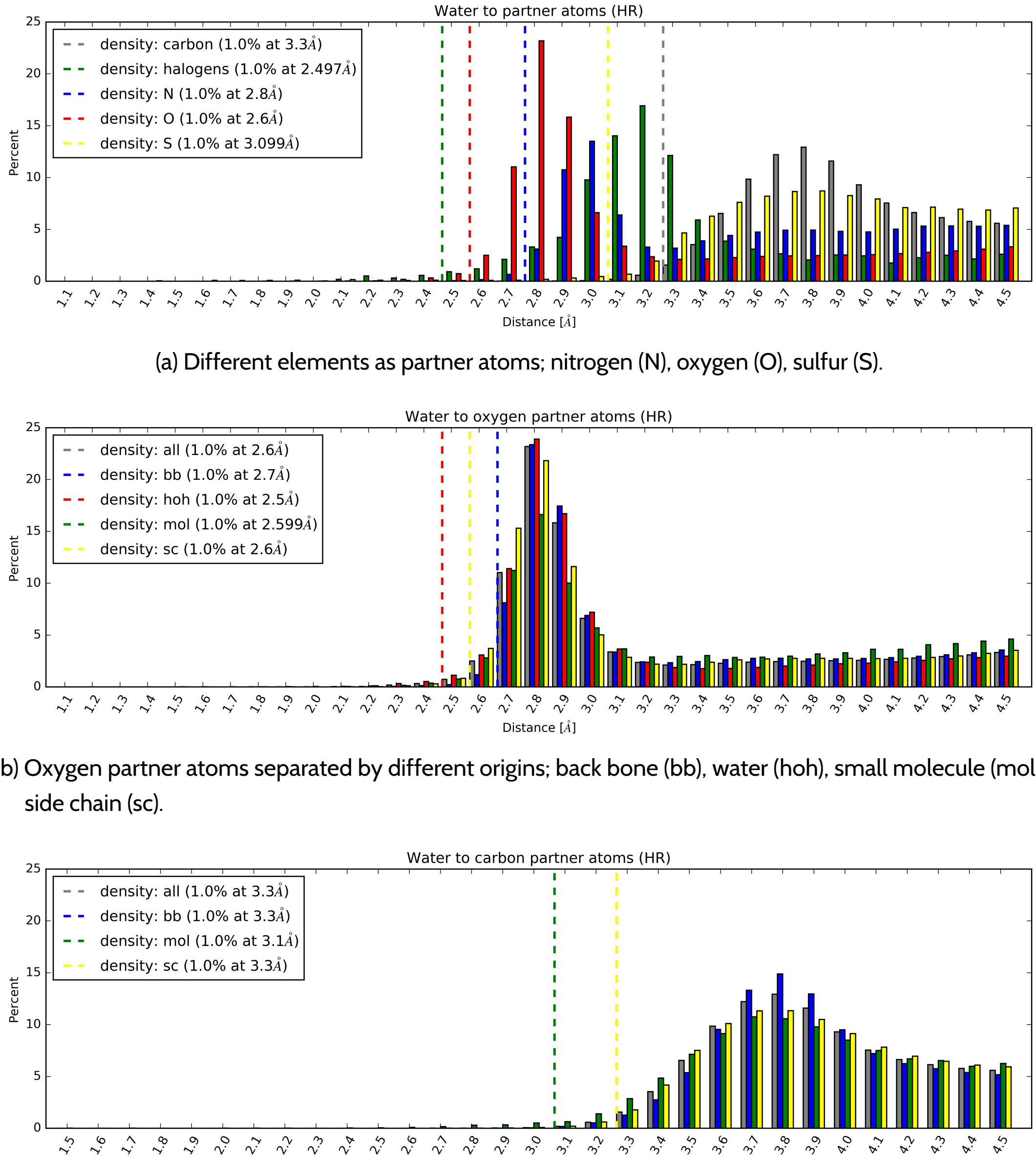
As expected, distances to polar atoms (nitrogen and oxygen) were shorter than the sum of the van der Waals radii (Table 3.1 and Figure 3.3a). Herein, oxygen atoms showed closer distances than nitrogen atoms, with the majority of surrounding atoms at 2.8 Å and 3.0 Å respectively. Therefore, the distribution of oxygen atoms was analyzed more closely. Oxygen atoms were classified according to their origin (Figure 3.3b). However, no significant difference in the distributions could be observed. The intensity of the peak resulting from small molecule oxygen atoms is lower than the remaining ones. This might be due to a higher amount of oxygen atoms in small molecules as shown by a higher percentage of oxygen atoms at larger distances.

Table 3.1: Van der Waals radii (r_{vdW}) of analyzed atoms.

Atom	r_{vdW}(atom) [Å]	r_{vdW}(atom + water[†]) [Å]
O	1.52	2.92
N	1.55	2.95
C	1.70	3.10
S	1.80	3.20
F	1.47	2.87
Cl[*]	1.75	3.15
Br	1.85	3.25
I	1.98	3.38

[†] Water radius was set to 1.4Å.

[*] Major contributor to the distance distribution in Figure 3.3a.



(a) Different elements as partner atoms; nitrogen (N), oxygen (O), sulfur (S).

(b) Oxygen partner atoms separated by different origins; back bone (bb), water (hoh), small molecule (mol), side chain (sc).

(c) Carbon partner atoms separated by different origins; back bone (bb), small molecule (mol), side chain (sc).

Figure 3.3: Volume normalized distance distributions of different atoms around a water molecule as the central group based on the high-resolution PDB subset from D2.

Weak hydrogen bonds, such as CH···O or halogen bonds, have been subject to several studies throughout the last years.[271–273] Therefore, their distances to water molecules were analyzed. Halogen atoms showed distances closer than the sum of the van der Waals radii (Table 3.1 and Figure 3.3a). The major contributor to the distance distribution was chlorine (80%). The main peak is just above the sum of the van der Waals radii of chlorine and water (peak: 3.2 Å; r$_{vdW}$: 3.15 Å). A visual inspection of 10% of chlorine atoms closer than 3.0 Å revealed only chloride ions and no small molecules containing a halogen. Sulfur and carbon atom distributions had significant intensities only at distances greater than the the sum of their van der Waals radii. Still, we took a closer look at carbon atoms to monitor any close distances that could arise due to so-called weak hydrogen bonds.[271–273] Independent of its origin, no significant amount of distances closer than the sum of the van der Waals radii could be observed (Figure 3.3c).

The information about water-surrounding distances was used in the water placement procedure. Herein, available areas within the protein have to be identified and atoms that could be closer to a water molecule than the sum of their van der Waals radii need to be respected adequately. For more details on the usage of the distance information see Chapter 4 and P1.

4
Placement and Scoring of Water Molecules

The previous two chapters display a requisite for placing water molecules (Chapter 3) as well as a means of validating the developed water placement procedure (Chapter 2). The derived interaction geometries were exploited to identify space within the protein complex suitable for water molecules. Due to a discretization of the interaction geometry surfaces, a clustering approach was used to derive final water positions. Subsequently, the derived water positions were optimized and subject to HYDE scoring (HYDE$_{water}$) to predict their energetic contributions (Figure 4.1). For a more detailed explanation of the method and validation procedure see P1. In this chapter, the identification of free space, the water placement procedure, and the estimation of water energy contributions are described.

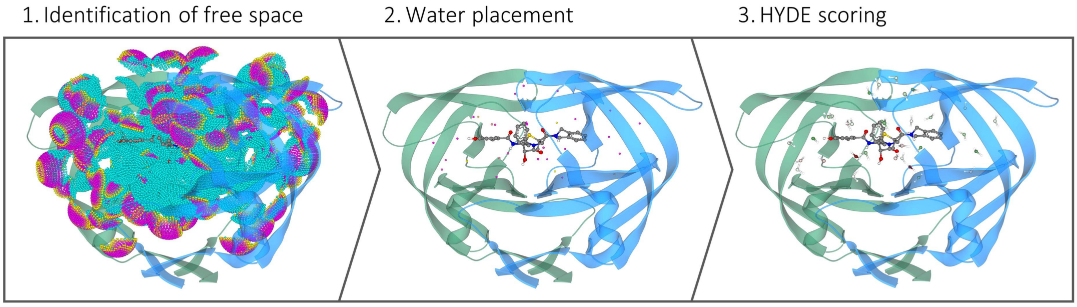
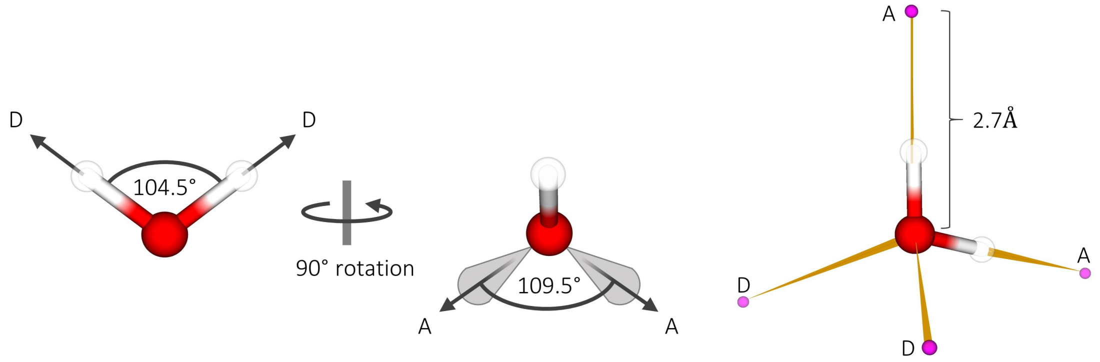


Figure 4.1: Overview of water placement and scoring work-flow; Protein structure: HIV protease (PDBid 1kzk[274]).

4.1 Identification of Water Molecules

Especially when it comes to cavities in protein structures or to protein-ligand binding sites, an implicit representation of water molecules may not be sufficient. On the one hand, the number of donor and acceptor functions needs to be considered. One water molecule has two donor and two acceptor functions in a tetrahedral arrangement, thus a maximum of two acceptors and two donors can be satisfied by one water molecule (Figure 4.2). On the other hand, and directly in connection with the previous aspect, the orientation of the water molecule is important. Especially if multiple atoms with donor and acceptor functions are in the vicinity of one water molecule they might not be arranged such that the water molecule can satisfy all hydrogen bond functions (Figure 4.2b).



(a) Geometric arrangement of acceptor (A) and donor (D) functions of a water molecule.

(b) Tetrahedral arrangement of donor (D) and acceptor (A) atoms around a water molecule in ideal hydrogen bond distance (2.7Å).

Figure 4.2: Geometric specifications of water molecules.

The HYDE scoring function played an essential role in the development of the water placement algorithm. HYDE is sensitive to geometric criteria of hydrogen bonds. Thus, explicit water molecules are mandatory. However, water molecules are not always resolved, especially in X-ray structures with less than 2.5 Å resolution. On the other hand, scoring functions are applied to protein-ligand docking poses where water molecules are not available.

For the above reasons, a consistent availability and representation of water molecules in protein structures is needed. The developed water placement procedure consists of two steps (Figure 4.3): (1) The identification of free space within protein structures and (2) the generation of explicit water positions.

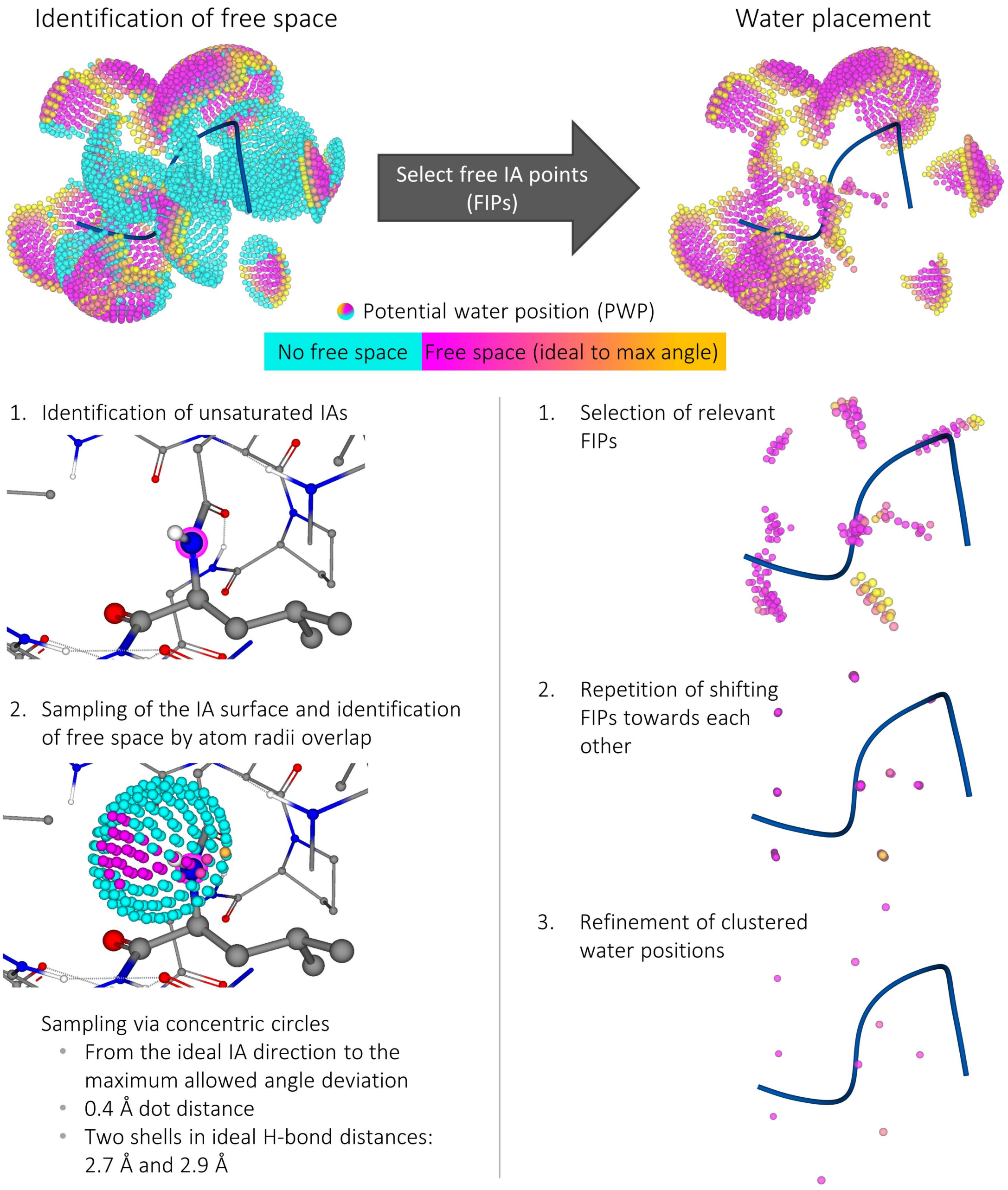

Identification of free space
Water placement
Select free IA points (FIPs)
Potential water position (PWP)
No free space
Free space (ideal to max angle)
1. Identification of unsaturated IAs
2. Sampling of the IA surface and identification of free space by atom radii overlap
Sampling via concentric circles
From the ideal IA direction to the maximum allowed angle deviation
0.4 Å dot distance
Two shells in ideal H-bond distances: 2.7 Å and 2.9 Å
1. Selection of relevant FIPs
2. Repetition of shifting FIPs towards each other
3. Refinement of clustered water positions

Figure 4.3: Overview of free space identification (left) and water placement procedure (right). The identification of free space has two main steps and forms the basis for the water placement, which consists of three main steps. IA = interaction; H-bond = hydrogen bond; FIPs = free interaction points.

4.1.1 FSI – Free Space Identification

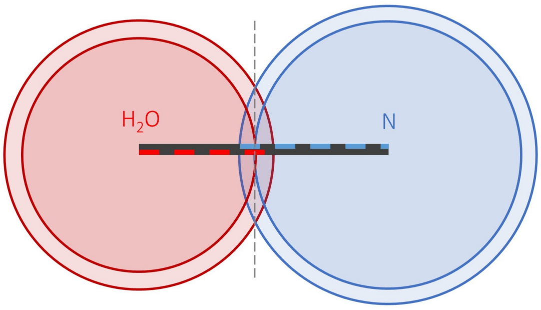
The identification of free space (FSI) is based on derived interaction geometries from a large-scale evaluation of protein structures (Chapter 3 and D4). Using a discretization approach, explicit positions for water molecules were generated, further called potential water positions (PWP). Herein, the interaction geometries were discretized using concentric circles (Figure 4.3 *Identification of free space*). Depending on the type of interaction together with its chemistry type and its geometry type, different geometric criteria, i.e. cone or spherical rectangle, as well as angle specifications, were applied (Appendix B.2).

Concerning the discretization, two different aspects have to be considered – the accuracy as well as the run-time. Both aspects mutually influence each other: the smaller the distance between PWPs, also called dot distance, the larger the run-time, while more coarse-grained sampling reduces the run-time, but leads to a loss in accuracy.

The dot distance is directly connected to the accuracy of the method. Therefore, a criterion, further called overlap criterion, for 'available' versus 'occupied' needs to be determined before evaluating the accuracy dependence of the dot distance. Every PWP was subject to the overlap criterion. Herein, the distances to atoms surrounding the PWP were evaluated. In case of polar atoms, adjusted atom radii were used (Equation 4.2). Due to potential hydrogen bonds from a water molecule to a polar atom, the distances can be shorter than the sum of their van der Waals radii (Figure 3.3). Thus, the radii of polar atoms were adjusted to allow potential hydrogen bonds (Figure 4.4). Since all analyzed apolar atoms showed no van der Waals radii overlap (Figure 3.3), their van der Waals radius was used without any adjustment for the determination of accessibility of a PWP.

$$2.6\text{Å} = 0.895 \cdot r_{vdW}\left(a_{polar}\right) + 0.866 \cdot r_{H_2O} \tag{4.1}$$

$$\implies r_{adj}\left(a_{polar}\right) = 0.895 \cdot r_{vdW}\left(a_{polar}\right) - 0.134 \cdot r_{H_2O} \tag{4.2}$$



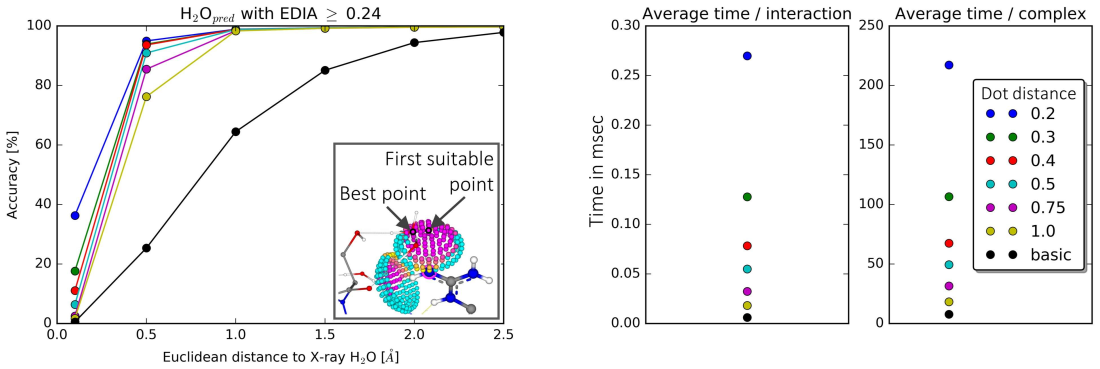
Black line = min distance water---N = 2.6 Å
Red dashed line = r(H_2O) = 1.4 Å
Blue dashed line = r_{vdW}(N) = 1.55 Å

Red outer circle = water surface with r(H_2O)
Red inner circle = 86.6 % r(H_2O)
Blue outer circle = nitrogen surface with r_{vdW}(N)
Blue inner circle = 89.5 % r_{vdW}(N)

Figure 4.4: Adjustment of polar atom radii according to water-surrounding analysis (Section 3.3).

After defining the overlap criterion, the best dot distance, i.e. the distance between individual PWPs, was evaluated. Different dot distances were examined for their achieved accuracies to identify an optimal balance between accuracy and run-time. Herein, accuracy is defined as an available PWP (further called free interaction point or FIP) close to a crystallographically observed water molecule.

The evaluation was based on the previously compiled high-resolution PDB subset (Chapter 2.2). Finally, a dot distance of 0.4 Å led to a high accuracy, comparable to accuracies achieved with smaller dot distances (Figure 4.5a), while the run-time was kept at an intermediate level (Figure 4.5b). The developed FSI was compared to a basic procedure, in which only the ideal interaction direction was used to identify free space. This procedure is faster, due to no sampling, but the accuracy is significantly lower (Figure 4.5).

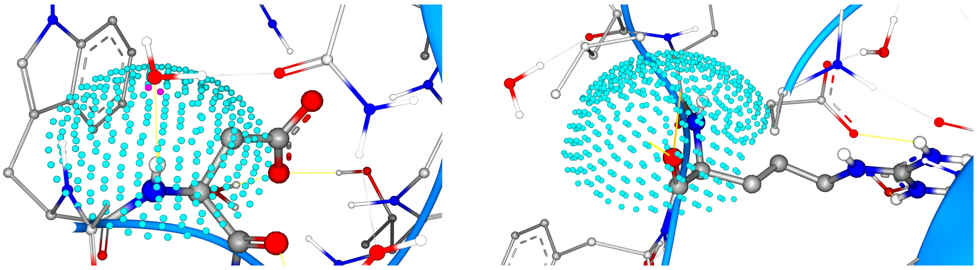


(a) Achieved accuracies by different voxel distances. Lines connecting the individual points are for visualization purpose only.

(b) Run-time differences by voxel distances for single interactions and whole complexes.

Figure 4.5: Evaluation of different voxel distances for the discretization of interaction geometries; Best PWP is used to evaluate the accuracy of the free space identification (figure inlet); 'Basic' is an approach where only the ideal interaction direction is used for the identification of free space.

Overall, the FSI showed a high accuracy in correctly identifying free space in protein structures for water molecules (Figure 4.6). Thus, the method can be applied to both, implicit handling of water molecules for areas where the exact orientation is less relevant, i.e. at the protein surface, but also as a starting point for explicitly placing water molecules.



(a) Accurate identification of a confined area for a water molecule (Asp43, PDBid 1isp[7]).

(b) Accurate identification of no free space for a water molecule (Arg112, PDBid 3inh[275]).

Figure 4.6: Examples for the identification of free space in protein structures; purple spheres = free space; light blue spheres = occupied space.

4.1.2 WarPP – Water Placement Procedure

The basis for the water placement procedure (WarPP) are unsaturated interaction directions of protein or ligand atoms. Since protein structures usually do not contain hydrogen atoms, we used Protoss[276] for the optimization of the hydrogen bond network. Without considering the crystallographically observed water molecules, hydrogen atoms were added and the hydrogen bond network was optimized as well as protomers and tautomeric states adjusted.[276]

Clustering of PWPs

Unsaturated interaction directions, as well as directions with non-ideal geometries, were selected for subsequent water placement using the FSI (Figure 4.3 *Water Placement*). For the water placement, rotatable functional groups were modeled as capped cones (Figure 4.7). Modeling explicit interaction directions (Figures 4.7a and 4.7b) had the disadvantage of regions with no FIPs at all as well as differing geometric scores, which depend on the initial Protoss run and influence the water placement. In contrast, a capped cone (Figure 4.7c) had a continuous distribution of FIPs with the same geometric score on the same concentric circle.

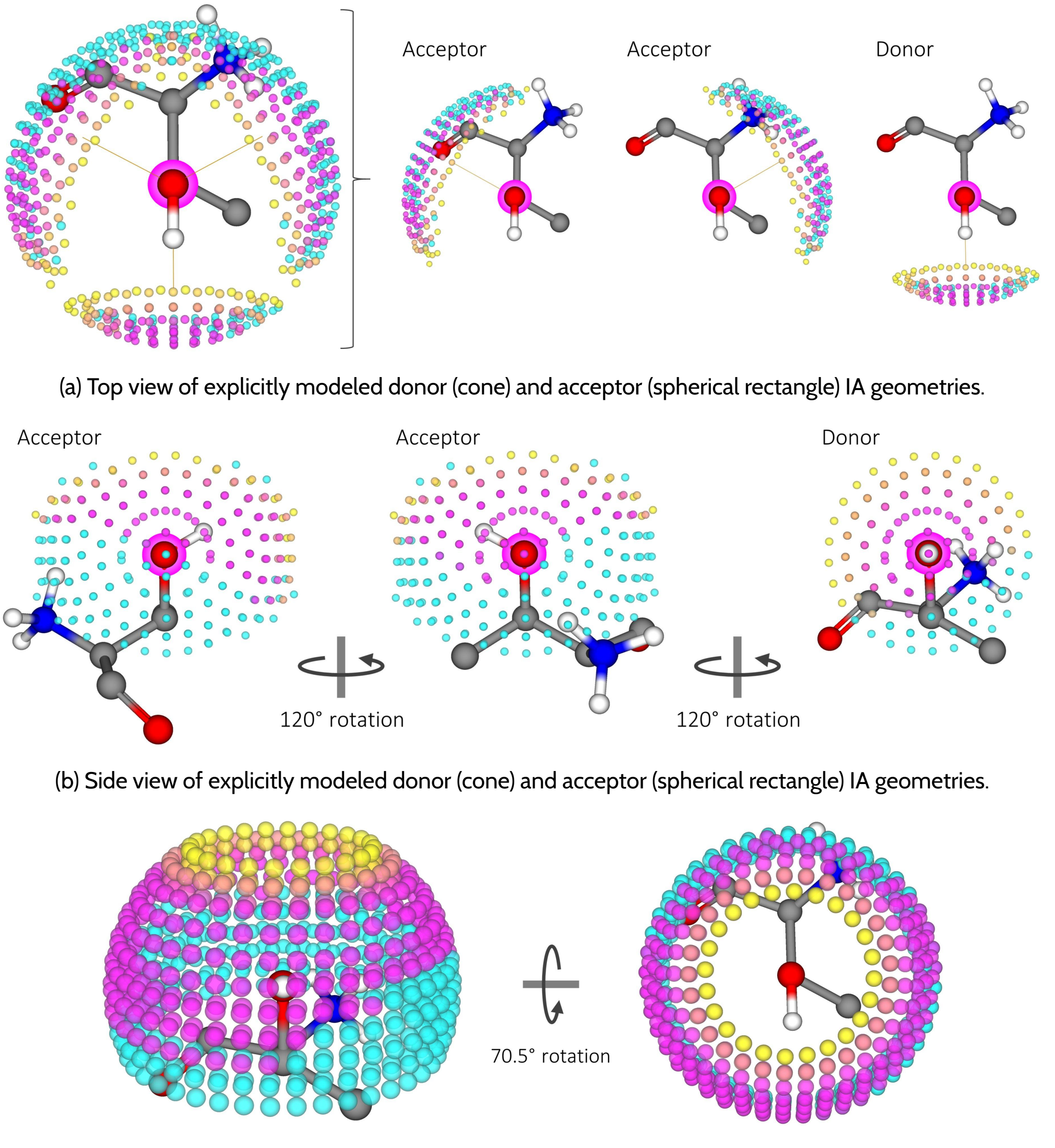
Every FIP displayed a potential position of an explicit water molecule. However, a selection had to be made to place a biologically reasonable amount of water molecules, ideally the same amount as crystallographically observable. Diverse clustering approaches exist that are applied to drug discovery problems.[277] However, several aspects had to be included during clustering, such as the geometric score, the origin of the FIP, the hydrogen bond angles as well as distances. In order to include all of them, we developed our own approach. An approach was evaluated, where the points would cluster themselves, i.e. the points are shifted towards each other until they have approximately the same position. The approach is agglomerative, which means that single FIPs are merged to receive final water positions. More details on the water placement procedure and its parametrization can be found in P1.

Optimization of Water Positions

During the water placement procedure, FIPs were drawn towards better hydrogen bond geometries. Due to the maximum cluster distance of 1.4 Å water positions can be closer to each other than the allowed minimum hydrogen bond distance. Therefore, a refinement step was added to ensure correct distances between the placed water molecules.

A gradient-based optimization procedure was selected with four optimization criteria: (1) Optimization of water-water distances; (2) Maintaining hydrogen bonds to polar surrounding atoms of water molecules; (3) Avoiding clashes with surrounding atoms; (4) Maintaining hydrogen bonds between the placed water molecule and their atoms of origin, i.e. those atoms that provided FIPs for the water placement.

The numerical optimization was performed using an in-house implementation of the BFGS algorithm (see Nocedal *et al.*[278] for a detailed description of the BFGS algorithm).



(a) Top view of explicitly modeled donor (cone) and acceptor (spherical rectangle) IA geometries.

(b) Side view of explicitly modeled donor (cone) and acceptor (spherical rectangle) IA geometries.

(c) Implicit representation of IA geometries using a capped cone.

Figure 4.7: Interaction (IA) geometries for the hydroxyl oxygen atom of a threonine side chain; For color coding of IA surface points see Figure 4.3.

The developed optimization strategy was evaluated using crystallographically observed water positions. Based on the previously compiled high-resolution PDB subset (Chapter 2.2) all protein-ligand complexes with well resolved ligands (EDIA$_m$ ≥ 0.8) were selected. For the evaluation, only structurally relevant water molecules, i.e. with an EDIA between 0.24 and 3.5 as well as two hydrogen bonds to ligand or protein atoms of the active site, were used. Overall, about 20,000 water molecules were used for the evaluation. The selected crystallographically observed water molecules were optimized. 85% of the water molecules were shifted less than 0.5 Å away from its crystallographic position, 96% less than 0.75 Å and 97% less than 1.0 Å (Figure 4.8a). Thus, we were confident to use the developed optimization strategy to refine placed water positions.

Evaluation and Conclusion

Four different criteria were analyzed for the evaluation of WarPP: (1) Sensitivity, (2) precision, (3) water–water distance distribution, and (4) EDIA values for placed water molecules.

For the evaluation of placed water molecules, we need to differentiate between sensitivity and precision (Table 4.1). The sensitivity (Equation 4.3) gives information about how many crystallographically observed water molecules are matched by placed water molecules, while the precision (Equation 4.4) contains information about how many water molecules are placed that are not matched by crystallographically observed water molecules.

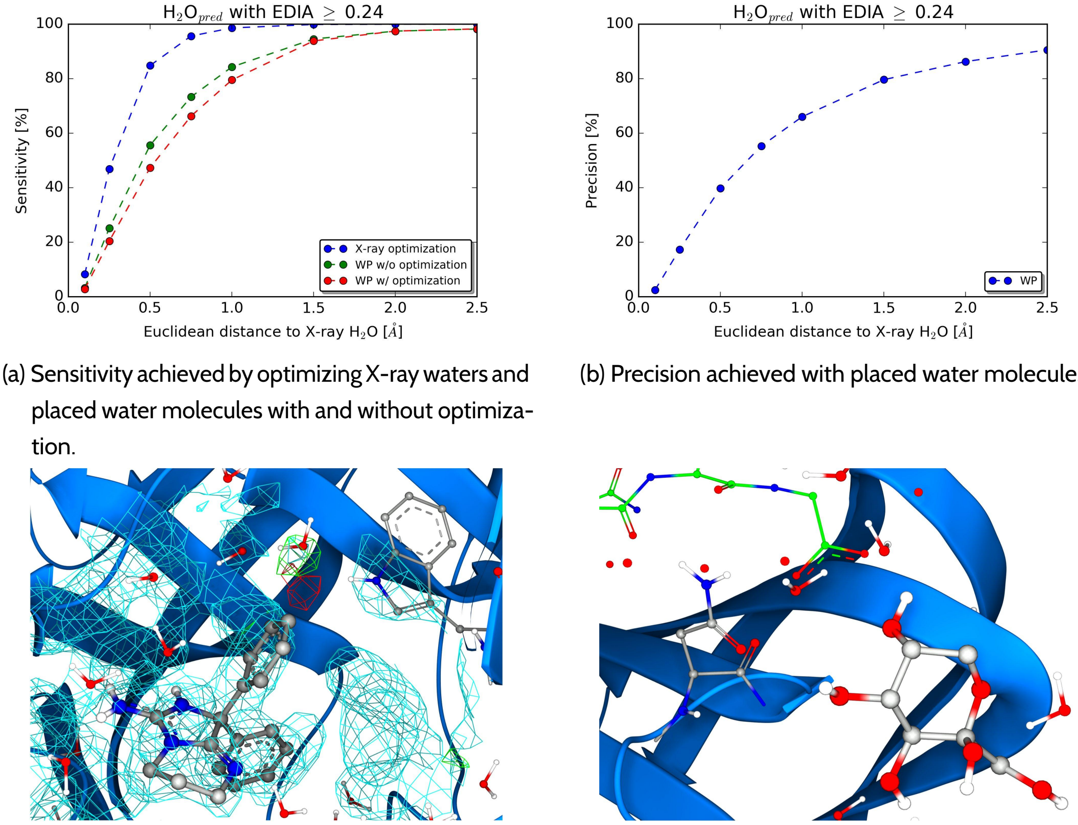
$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{TP}{\text{Number of X-ray waters}} \tag{4.3}$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{TP}{\text{Number of placed waters}} \tag{4.4}$$

Table 4.1: Terminology of a confusion matrix for water placement; X Å = selected distance.

		Observed category	
		X-ray water available	No X-ray water
Predicted category	Placed	*True positives* (TP) = placed water molecule in X Å distance to a crystallographically determined water molecule	*False positives* (FP) = placed water molecule without a crystallographically determined water molecule within X Å distance
	Not placed	*False negatives* (FN) = crystallographically determined water molecule without a placed water molecule within X Å distance	*True negatives* (TN) = no crystallographically determined water molecule and no placed water molecule within X Å distance (cannot be calculated)

WarPP was evaluated using the same data set as for the evaluation of the optimization strategy. The sensitivity of our method was calculated using placed water molecules and their distance to the crystallographically observed water molecules. We achieved a sensitivity of 47% within 0.50 Å and of 80% within 1.0 Å (Figure 4.8a). The precision was calculated using all water molecules with a sufficient EDIA, but it was not limited to water molecules with two interactions. Thus, our method achieved a precision of 40% within 0.5 Å and of 66% within 1.0 Å (Figure 4.8b). The drop in precision compared to sensitivity has multiple sources. First, water molecules might be too flexible to be defined to a specific position, i.e. surface water molecules or those at the rim of the active site. Second, water molecules might not have been modeled in the crystal structure due to different reasons (Figure 4.8c). And third, water molecules might be part of another crystal symmetry, which was not considered during the evaluation process (Figure 4.8d).



(a) Sensitivity achieved by optimizing X-ray waters and placed water molecules with and without optimization.

(b) Precision achieved with placed water molecules.

(c) Unmodeled water molecules within an active site (PDBid 3igb[275]); Blue mesh = electron density map (2fo-fc map) at 1σ; red/green mesh = electron density difference map (fo-fc map) at -3σ and +3σ, respectively.

(d) Water molecules placed at positions of polar atoms of another asymmetric unit (PDBid 1ovp[279]); Red spheres and ball-and-stick representation with green carbons = neighboring crystal contact.

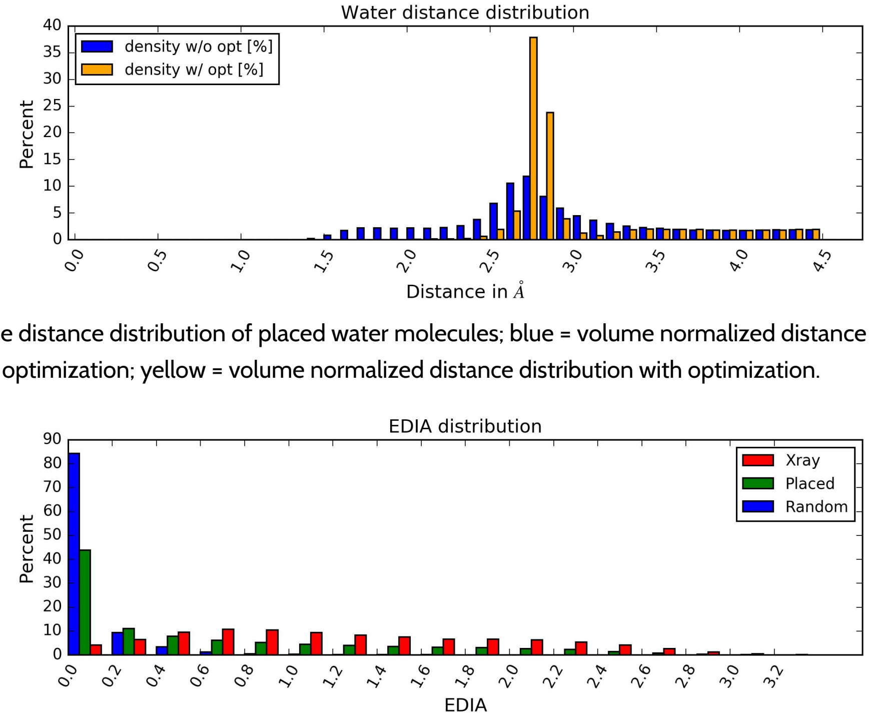
Figure 4.8: Sensitivity and precision of water placement procedure with examples cases; Water molecules in ball-and-stick representation = placed water molecules. Dashed lines in a) and b) are for visualization purpose only.

In addition to the sensitivity and precision, the pair-wise distances of the placed waters were evaluated. Placing water molecules should lead to correct hydrogen bond distances between the water molecules. Thus, measuring water···water distances contains useful additional information about the hydrogen bond network beside the absolute number of placed water molecules. Herein, placed water molecules are expected to show a similar distance distribution to crystallographically observed water molecules. Compared to observed water distance distributions (Figure 3.3b), the placed water molecules show a narrower peak with a maximum between 2.7 Å and 2.8 Å (Figure 4.9a). A drop in density can be observed at 3.1 Å, which is due to the function used for the optimization of water molecules. Overall, the distance distributions between observed and placed water molecules agree well with each other.

The electron density was exploited to further validate the quality of placed water molecules. Herein, the EDIA values for crystallographically observed, placed, and randomly distributed water molecules were calculated (Figure 4.9b). For the random distribution of water molecules, the binding site was evenly distributed using a 3D grid with a voxel distance of 2.2 Å. The active site residues were used as maxima and minima for the grid definition. Every voxel within 8.0 Å of the ligand was assigned either available to water molecules or unavailable based on the previously defined overlap criterion. If the voxel was available, the EDIA for the respective position was calculated. The grid was shifted separately in x, y, and z direction by 1.1 Å to achieve a better sampling of the active site. The voxel distance should roughly resemble hydrogen bond distances. However, a distance of 2.6 Å let to almost no available voxels on the 3D grid. Thus, the distance was reduced to 2.2 Å. Compared to the EDIA distribution of crystallographically observed water molecules, placed water molecules have a larger amount of EDIA values between 0.0 and 0.2. 45% of the placed water molecules achieve EDIA values above 0.4 while only 1% of randomly distributed water molecules have an EDIA greater than 0.4. Overall, the placed water molecules have a better quality than randomly distributed water molecules but are less accurate than crystallographically observed ones.

No other available water placement methods were evaluated on a comparable sized data set. Often, a data set consisting of the same protein family was used and only conserved crystallographic water positions were considered for water placement predictions.[145,146,150] Additionally to the small data sets, the sensitivity is mainly calculated for placed water molecules within 1.5 Å or 2.0 Å distance of crystallographically observed ones. Compared to other methods, WarPP achieves a high sensitivity while keeping the precision at a fair level.

Compared to the FSI the sensitivity drops significantly at low distances due to the great change in the number of available positions considered for the evaluations, i.e. for the FSI the whole interaction surface is used, while for WarPP only one discrete position is considered. Additional problems for the water placement are different amino acid side chain orientations that cannot be differentiated based on the electron density. For example, histidine side chains can flip, resulting in the same electron density but in swapped polar and apolar atoms. Thus, in one state an interaction is feasible while in the other a carbon atom occupies the position. Crystal symmetry displays another interesting aspect for analysis. On the one hand, as described above, some of the placed water molecules are actually crystallographically available when symmetry operations are applied to the protein structure. At the same time, further available amino acid side chains from the symmetry unit could lead to a more precise placement due to their occupancy of space where, without considering crystal symmetry units, space for a water molecule would be available. However, especially the consideration of crystal symmetry units highly depends on the biological relevance of the crystal contact. If it actually has



(a) Pair-wise distance distribution of placed water molecules; blue = volume normalized distance distribution without optimization; yellow = volume normalized distance distribution with optimization.

(b) EDIA distribution of crystallographically observed water molecules (red), placed water molecules (green) and randomly placed 'water probes' (blue).

Figure 4.9: Evaluation of placed water molecules by pair-wise distances and EDIA.

a biological relevance, i.e. the functional protein is displayed by the crystal contact, it should be considered during water placement. If it is only a crystal artifact, i.e. the crystal contact only exists due to crystallization conditions, it should not be taken into account during water placement.

Runtime

The placement of water molecules in the active site (8.0 Å around the ligand atoms) takes on average 6.8 sec and 0.08 sec per water molecule placed.

For further methodical details and the evaluation of the water placement please refer to P1.

4.2 Energetic Contribution of Water Molecules

The energetic contribution of water molecules is of interest for various reasons: First, water molecules contribute to the overall binding affinity, not only of protein-ligand complexes, but also to protein-protein interactions or the stability of a protein itself. Second, replacing water molecules is a common strategy in drug development to enhance the binding affinity of a small molecule to its protein binding site. Therefore, the most promising water molecules need to be identified to apply rational-driven ligand alterations.

However, the most difficult aspect is the estimation of the energetic contribution of a single water molecule. As described in Chapter 1.1.1, it is experimentally hardly feasible to measure the contribution of a single water molecule. Even if a water molecule is displaced by a ligand, it is always replaced with an extension of the ligand. Thus, the energy difference is always a result of the water molecule displacement in combination with the ligand extension, which also contributes to the binding affinity. According to theoretical considerations (Chapter 1.1.3), the maximum free energy contribution of a single water molecule is estimated to be between -0.7 and -2 kcal mol^{-1}.

An aspect that is especially relevant for the integration of water molecules in a scoring function, is their representation. Water molecules in bulk form up to four hydrogen bonds with an average of two geometrically high quality hydrogen bonds.[280–283] Therefore, the energetic contributions of water molecules may need a different handling than other ligand atoms during scoring.

4.2.1 HYDE Scores for Water Molecules

Diverse aspects of the calculation of the HYDE scoring function[15–19] were adapted in the course of this dissertation. Below, only the main alterations that directly implicate water scores are described. For more information on the HYDE scoring function please refer to Appendix B.1. For differentiation reasons, the HYDE version[b] that built the foundation of this dissertation will be called HYDE$_{2012}$.

Calculation of ΔG$_{\text{HYDE}}$

The total HYDE score consists of a saturation and dehydration term.

$$\Delta G_{\text{HYDE}} = \sum_{\text{atoms i}} \Delta G^{i}_{\text{saturation}} + \Delta G^{i}_{\text{dehydration}} \tag{4.5}$$

The dehydration term is calculated differently for polar and apolar atoms. While for polar atoms their hydrogen bond functions j (HBj) are considered, the difference in accessible surface area (Δacc^{i}) is

[b] For further information please refer to the dissertation of N. Schneider.[17]

used for apolar atoms.

$$\Delta G^{i}_{\text{dehydration}} = \Delta G^{i,polar}_{\text{dehydration}} + \Delta G^{i,apolar}_{\text{dehydration}} \tag{4.6}$$

$$\Delta G^{i,polar}_{\text{dehydration}} = -2.3 \cdot RT \cdot p\log P^{i} \cdot \sum_{\text{HB j}} w^{j} \cdot p^{j}_{dehyd} \tag{4.7}$$

$$\Delta G^{i,apolar}_{\text{dehydration}} = -2.3 \cdot RT \cdot p\log P^{i} \cdot \Delta acc^{i} \tag{4.8}$$

R is the gas constant, T the temperature, and $p\log P$ the partial $\log P$ parameter based on octanol-water partition coefficients (Appendix B.1.1). For polar atoms, the contribution of hydrogen bond functions is multiplied by a weighting term w^{j} (Equation 4.13). If an atom has only one hydrogen bond function, the weighting term w^{j} is one. In case an atom can form multiple hydrogen bonds, the weights are distributed among the different hydrogen bond functions. While the probability of dehydration p^{j}_{dehyd} (Equation 4.12) is considered for the dehydration term, the geometric quality f^{j}_{dev} of the hydrogen bond influences the saturation term. The geometric quality f^{j}_{dev} of a hydrogen bond is determined by four measurements: heavy atom distance, head-head distance, i.e. of the electron pair and the hydrogen atom, donor, and acceptor angles (see P1 for more information). More details about the assignment of hydrogen bond weights and the dehydration probability are given later in this chapter.

$$\Delta G^{i}_{\text{saturation}} = \frac{2.3 \cdot RT}{F_{sat}} \cdot p\log P^{i} \cdot \sum_{\text{HB j}} w^{j} \cdot f^{j}_{dev} \tag{4.9}$$

The principle calculations of the HYDE scoring function (Equations 4.5 – 4.9) were not altered. However, the underlying calculations were changed with a special focus on two aspects: (1) A more accurate representation of water molecules and (2) a reduction of discrete decisions. The latter point is of importance for the geometric optimization of protein-ligand structures, called GeoHYDE. Discrete decisions lead to steps in functions, which are difficult for optimization algorithms.

The HYDE score represents the difference between the unbound and the bound state of a protein-ligand interaction (Figure 4.10). In order to differentiate between hydrogen bonds already existing in the unbound state, a classification was introduced in HYDE$_{2012}$: external IA (interaction), internal IA, no IA, or covered. *External IAs* were those that exist only in the bound state, while *internal IAs* already existed in the unbound state. *No IAs* were hydrogen bond functions that were either fully or partially accessible to water molecules (Section 4.2.1). While *covered* meant that the hydrogen bond function was not accessible to water and does not participate in a hydrogen bond.

In order to remove these classifications and consider the difference between the unbound and bound state directly, we calculated the HYDE score for both states – unbound and bound – separately (Equation 4.10).

$$\Delta G_{\text{HYDE}} = \Delta G_{\text{HYDE}}\,(\text{bound}) - \Delta G_{\text{HYDE}}\,(\text{unbound}) \tag{4.10}$$

A change in the HYDE score calculation was only necessary for the hydrophilic score contributions, since the hydrophobic term of the HYDE score already included the difference of the surface area accessibility between the unbound and bound state. The change in the calculation of the hydrophilic score contribution allowed the inclusion of the quality of protein···water and ligand···water interaction in more detail. In the exemplary protein-ligand binding site in Figure 4.10, red hydrogen bonds exist in the unbound as well as the bound state. Thus, they cancel each other out and do not contribute to the overall binding affinity. If water molecules are present that do not change upon binding, such as enclosed ones, they do not contribute to the binding affinity. Yellow depicted hydrogen bonds are those to water, which change upon binding and green are those hydrogen bonds that are formed during binding. The geometric scores for the yellow depicted hydrogen bonds are approximated by implicit water molecules (Section 4.2.1 *Inclusion of Implicit Water Molecules*). Thus, the difference between the hydrogen bond in the unbound and the bound state can lead to an affinity gain, in case the hydrogen bond in the unbound state was qualitatively less ideal, or to a drop in affinity, in case the hydrogen bond is more restricted in the bound state.

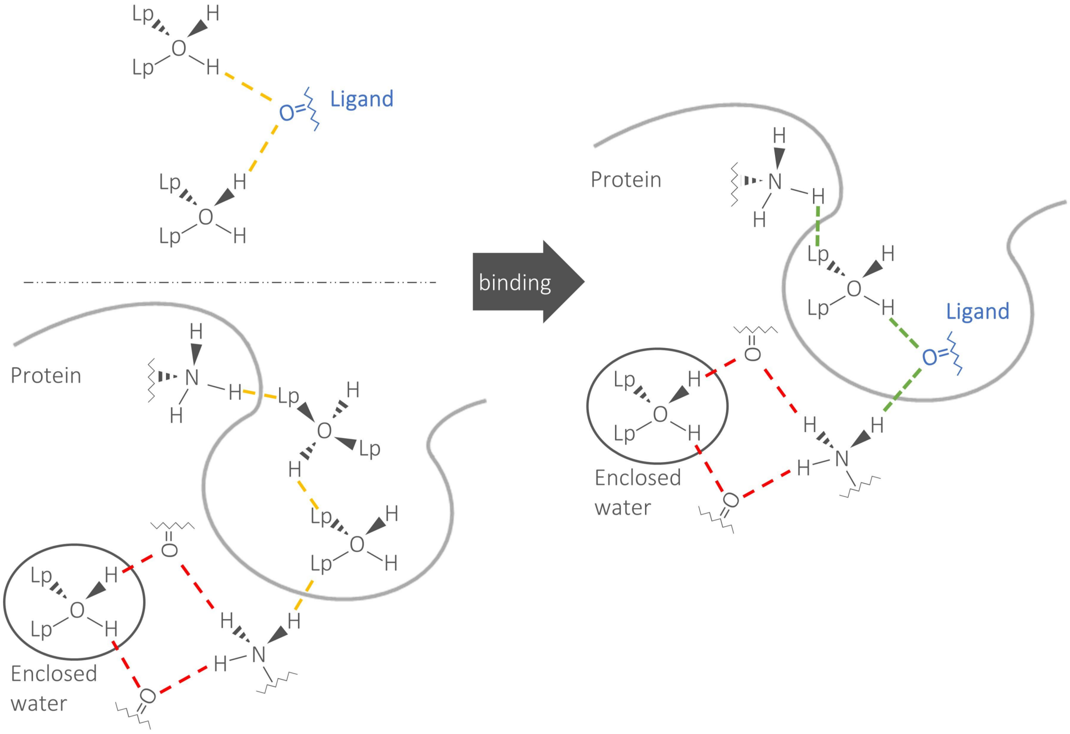
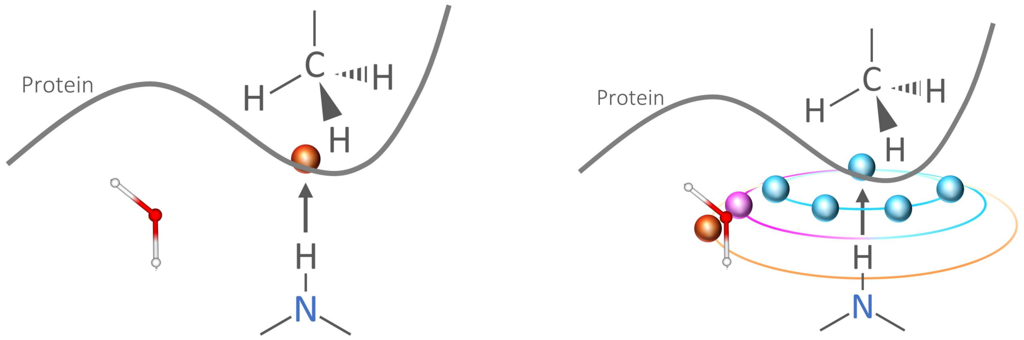


Figure 4.10: Calculation of the total HYDE score by explicitly calculating the unbound and bound HYDE score contributions; red dashed lines = hydrogen bonds (HB) that do not change between unbound and bound state; yellow dashed lines = HB to water molecules in the unbound state; green dashed lines = inter HB formed upon protein-ligand binding.

Calculation of the Dehydration Probability

The basic algorithm used for calculating the dehydration probability, the probability if a hydrogen bond function is accessible to water or not, was based on a search in ideal interaction direction in HYDE$_{2012}$ (Figure 4.11a). The dehydration probability was calculated as a function of the water volume that would fit in ideal interaction direction. If less than half a water volume would fit, the hydrogen bond function would be termed inaccessible ($p_{dehyd}(0.5V_{H_2O})$ = 0.0). From half a water volume to a full water volume the dehydration probability increased linearly ($p_{dehyd}(1.0V_{H_2O})$ = 1.0). The intention of using half a water volume, was to identify areas with shifted water molecules, i.e. those not in ideal interaction direction. However, half a water volume can lead to false positives, i.e. the identification as accessible, even though a water molecule would not fit, as well as false negatives, i.e. not enough space in ideal direction but enough space for a full water molecule located at a larger deviation angle.

Herein, the free space identification, introduced in the previous section, was applied (Section 4.1.1) to correctly identify those interaction directions, where no water molecule would fit, and those, where the placement of a water molecule would still be possible (Figure 4.11b).



(a) Basic algorithm: evaluation of free space in ideal interaction direction by testing for $0.5V_{H_2O}$.

(b) Sampling algorithm: searching the full interaction area with a full water molecule (see also Figure 4.3).

Figure 4.11: Basic and enhanced algorithm for the identification of free space in HYDE and the calculation of the dehydration probability p$_{dehyd}$.

The dehydration probability is now given by the geometric quality a water molecule would get at the first available PWP (Equations 4.11 and 4.12).

$$f_{dev}^{FSI} = f_{dev}(PWP) \tag{4.11}$$

$$p_{dehyd} = 1 - f_{dev}^{FSI} \tag{4.12}$$

Contribution of Hydrogen Bonds

Based on a study of small molecules, molecular surface area and experimental octanol/water partition coefficients – $\log P$ – were related.[18] The study showed that the hydrophilicity of small molecules is independent of the number of hydrogen bond functions of one polar atom, i.e. primary, secondary, and tertiary amines let to the same contribution to the $\log P$ value. These results led to the assumption that only the first hydrogen bond should contribute to the binding affinity, which is in contrast to an often applied scheme, where the energetic contribution of hydrogen bonds is assumed to be additive.

These results were integrated into the HYDE$_{2012}$ scoring function. Thus, the first hydrogen bond of a polar atom contributed most to the free energy, while the following hydrogen bonds contributed to a smaller extend. This model is called '100-20-10' model, due to the weight of 100% for the first hydrogen bond, 20% for the second, and 10% for the third, while more hydrogen bonds do not contribute any further. The additional 20% and 10% cannot be explained by experimental data derived from the $\log P$ analysis. Thus, a simplification of the '100-20-10' model was generated, called '100-0-0' model, which represented the assumption of the original publication. Another equally valid assumption from the $\log P$ results would be an equal distribution of energetic contribution among the hydrogen bond functions. Since discrete assignments of unequal weights lead to cliffs in the scoring function, which is inappropriate from an optimization point of view, the weighting-scheme was adapted to consider the hydrogen bond quality f^j_{dev}. Thus, a change in the hydrogen bond quality during optimization leads to a smooth transition of the hydrogen bond weights w (Equation 4.13).

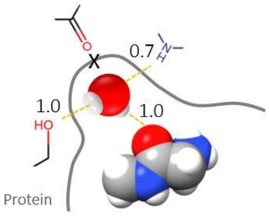
$$w_j = \begin{cases} 1 & \text{if \# IAs} = 1 \\ \frac{\left(f^j_{dev}\right)^2 + \left(\sum_{\text{IAs k}} 0.0001 \cdot p^k_{dehyd}\right) - 0.0001 \cdot p^j_{dehyd}}{\sum_{\text{IAs k}} \left(f^j_{dev}\right)^2 + (\#\text{IAs} - 1) \cdot \left(\sum_{\text{IAs k}} 0.0001 \cdot p^k_{dehyd}\right)} & \text{if \# IAs} > 1 \text{ and atom of origin} \neq \text{water} \\ \frac{\left(f^j_{dev}\right)^2 + \left(\left(\sum_{\text{IAs k}} p^k_{dehyd}\right) - p^j_{dehyd}\right) \cdot \frac{1}{16}}{\sum_{\text{IAs k}} \left(f^j_{dev}\right)^2 + (\#\text{IAs} - 1) \cdot \left(\sum_{\text{IAs k}} p^k_{dehyd}\right) \cdot \frac{1}{16}} & \text{if atom of origin = water} \end{cases} \tag{4.13}$$

For every hydrogen bond function j of an atom a the weight w_j is calculated under consideration of all hydrogen bond functions k of a. If a has exactly one hydrogen bond function (# IAs = 1), a weight of one is assigned. If atom a has multiple hydrogen bond functions (# IAs > 1), the weight is calculated according to equation 4.13. Herein, hydrogen bonds of the same quality f_{dev} contribute equally. If only one hydrogen bond is formed or one hydrogen bond has a significantly better geometry, this one contributes most. In case no hydrogen bond in formed, the dehydration penalty (p_{dehyd}, Equation 4.12) takes control of the weighting scheme, which ensures an energetic penalty for inaccessible and unsaturated hydrogen bond functions. Water molecules are treated separately from other protein or ligand atoms. Water molecules build on average two qualitatively ideal hydrogen bonds in bulk. Thus, the weighting scheme ensures a favorable energy contribution when the water molecule builds two geometrically high-quality hydrogen bonds in the protein-ligand complex. A comparison of the different weighting models is shown in Table 4.2.

Inclusion of Implicit Water Molecules

The dehydration probability was used to define the accessibility of hydrogen bond functions and can subsequently be used to determine the contributions of implicit water molecules.

Table 4.2: Weights for hydrogen bond functions for the central water molecule according to different weighting schemes; Geometric quality, f_{dev}, for each hydrogen bond (HB, yellow dotted line) is annotated at its respective interacting atom; X = covered hydrogen bond function (HBF).

		HB weights			
	Model	$HB_{H_2O \cdots Lig}$	$HB_{H_2O \cdots OH}$	$HB_{H_2O \cdots NH}$	$HBF_{H_2O \cdots X}$
	100-20-10	100%	20%	10%	0%
	100-0-0	100%	0%	0%	0%
	f_{dev}-based	37%	37%	20%	6%

Based on the HYDE theory, a stabilizing energy contribution is achieved due to a better geometric quality of hydrogen bonds in the bound state, than compared to hydrogen bonds to surrounding water molecules in the unbound state (F_{sat}, Equation 4.15).

$$F_{unsat}(T) = \frac{\Delta H_f + C_p(T - 273K)}{\Delta H_f + \Delta H_{273K-373K} + \Delta H_e} \tag{4.14}$$

$$F_{sat}(T) = 1 - F_{unsat}(T) \tag{4.15}$$

The enthalpy of fusion ΔH_f = 6.0 kJ mol^{-1} describes the necessary enthalpy to go from ice to liquid water at 273 K. C_p is the specific heat capacity of water (C_p = 0.0745 kJ mol^{-1} K^{-1}) and is constant between 273 K and 373 K. Thus, the enthalpy for heating water at a constant pressure of 1000 hPa can be calculated as $\Delta H_{273K-373K} = C_p \cdot \Delta T$ = 7.5 kJ mol^{-1}. The enthalpy of evaporation, to go from liquid water at 373 K to vapor, is ΔH_e = 40.7 kJ mol^{-1}. The fraction of unsaturated hydrogen bonds (F_{unsat}) can thus be calculated assuming that F_{unsat} in bulk water is proportional to the administered heat to the system.

F_{sat} represents the remaining fraction of satisfied hydrogen bonds, which is 0.855 at room temperature (298 K). Therefore, implicit water molecules can achieve at maximum a geometric quality of 0.85 (Equation 4.16) while their dehydration probability is set to 1.0 (Equation 4.17), which leads to a maximum energetic contribution of a hydrogen bond to an implicit water molecules of 0.0 kJ mol^{-1}.

$$f_{dev}^{implicit} = 0.85 \cdot (1 - p_{dehyd}) \tag{4.16}$$

$$p_{dehyd}^{implicit} = 1.0 \tag{4.17}$$

If a hydrogen bond is formed to an implicit water molecule in the unbound as well as bound state with exactly the same saturation and dehydration contribution, it will not affect the total binding affinity. In case a hydrogen bond is formed to an implicit water molecule in the unbound state, while upon binding a hydrogen bond between a protein atom and a ligand atom is formed, the energy difference contributes to the binding affinity. An energy gain is only achieved if the hydrogen bond between protein and ligand has a better quality than the implicit water hydrogen bond in the unbound state.

4.3 Implications of Placed Water Molecules on the HYDE Scoring Function

In $HYDE_{2012}$, explicit water molecules were counted as part of the protein. Thus, only the interactions between water and ligand were considered as contributing to the binding affinity. Through modeling implicit water molecules as well as considering the energy difference between the unbound and bound state, water molecules are now differentiated. If they do not change, i.e. if they are enclosed water molecules, they do not contribute to the overall energy difference. Explicit water molecules at the interface, i.e. protein-ligand interface, are considered implicitly in the unbound state and explicitly in the bound state. Thus, water molecules can now contribute to the overall binding affinity through water···protein and water···ligand interactions.

4.3.1 $HYDE_{water}$ – Scoring of Water Molecules

Different energetic contributions of water molecules are of interest for scoring. On the one side, individual water molecule scores contain information about the satisfaction of the water molecule itself. However, even if the water molecule itself contributes favorably to the overall binding affinity, the local surrounding might be unfavorable due to geometrically non-ideal arrangements. Therefore, apart from the contribution of a single water molecule also the so-called mapped affinities of amino acid and ligand atoms in close vicinity were evaluated (Figure 4.12).

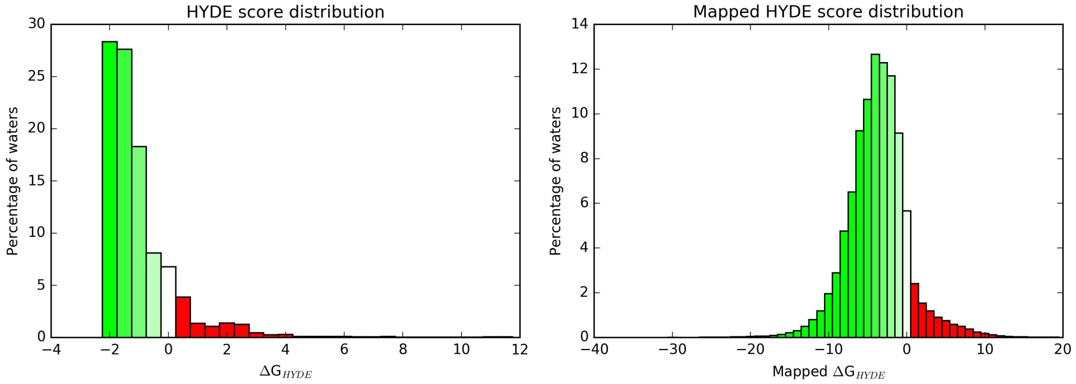
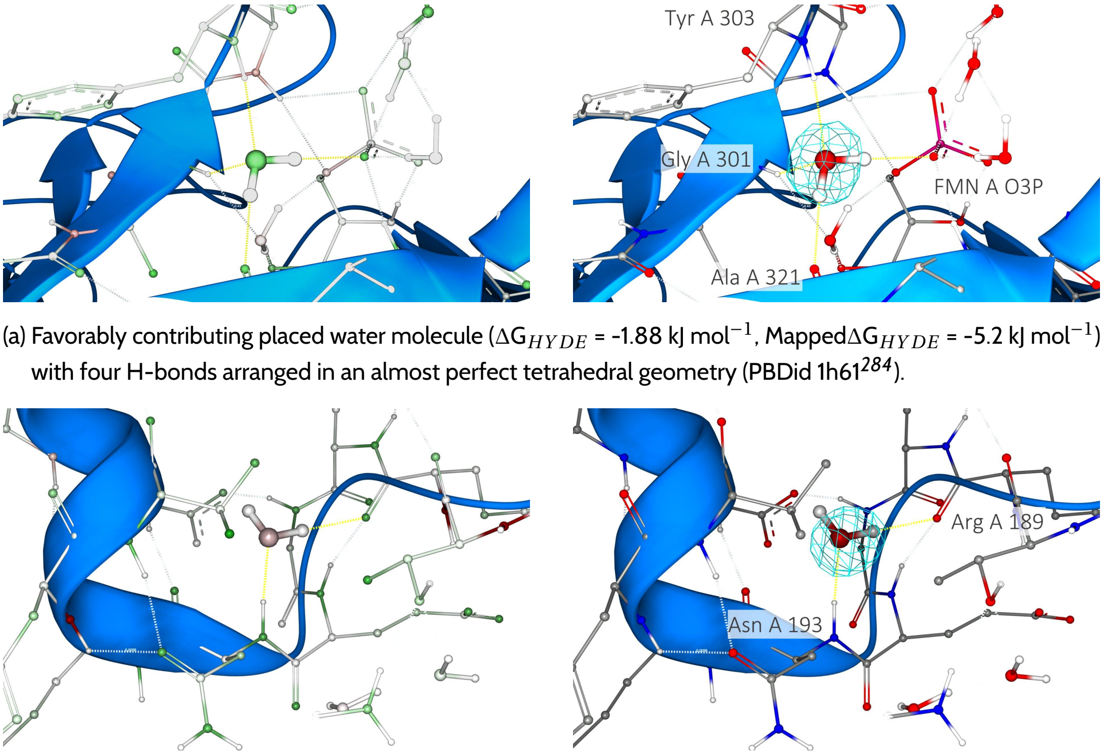


Figure 4.12: Single and mapped HYDE scores for placed water molecules; For each water molecule protein and ligand atoms within 8.0 Å surrounding were selected to form the water molecule's active site.

About 82% of all placed water molecules were predicted to be energetically favorable, 7% were energetically neutral, and 11% were unfavorably (Figure 4.12). Neutral water molecules are similar to bulk water. Most of those water molecules can be found at the outer rim of the active site, i.e. they interact with either protein or ligand or both and have free interaction directions that could interact with other surface water molecules. The distribution of mapped HYDE scores is similar to that of single water scores (85% favorably, 6% neutral, 9% unfavorably, Figure 4.12). The shape of the distribution is like a normal distribution with the main proportion around -5 kJ mol^{-1}.

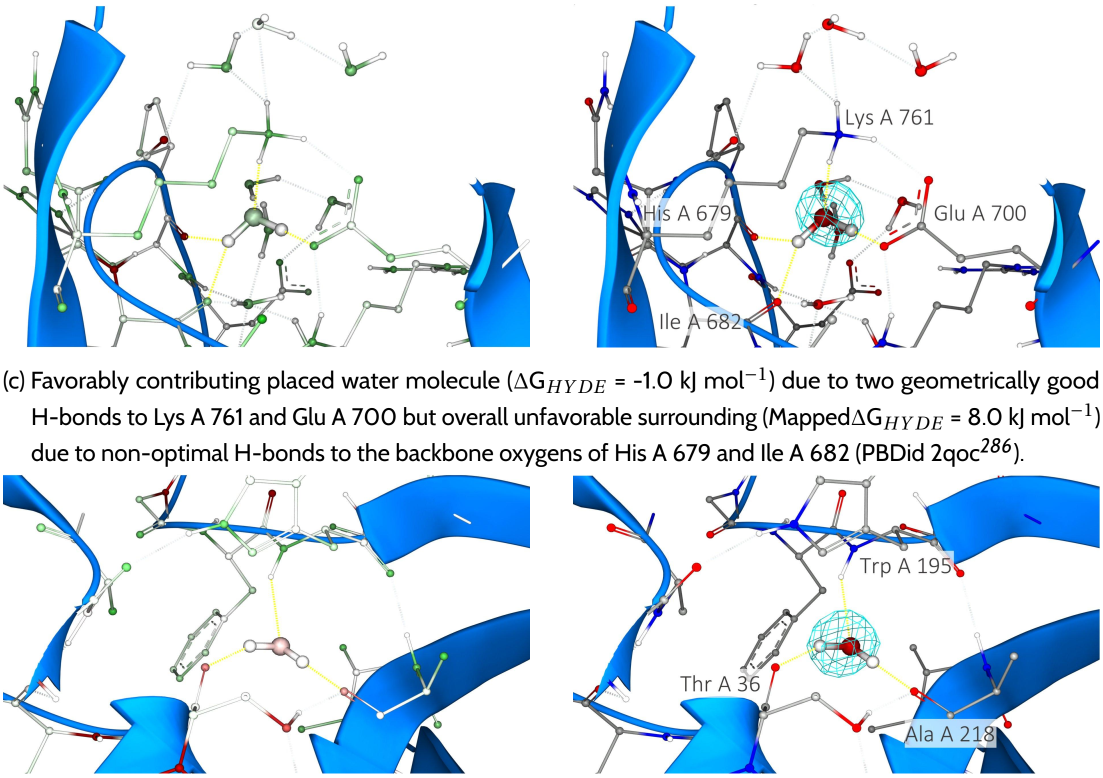
Examples for favorably and unfavorably contributing placed water molecules are given in Figures 4.13a and 4.13b, respectively. In both cases the energetic contribution of the surrounding is favorable (MappedΔG$_{HYDE}$ < 0 kJ mol^{-1}). In the first example (Figure 4.13a), the water molecule forms four geometrically ideal hydrogen bonds with its surrounding atoms. Thus, the water molecule itself as well as its the surrounding is satisfied. In the second example (Figure 4.13b), however, the water molecule participates in two hydrogen bonds: One geometrically ideal with the backbone nitrogen of Asn and one non-optimal to Arg. Since the oxygen atom of Arg participates in another hydrogen bond, which is geometrically ideal, it is energetically satisfied due to the larger contribution of the ideal hydrogen bond to the overall energy (Equation 4.13). The remaining hydrogen bond functions of the water molecule are covered due to a hydrophobic surrounding. Thus, the energy of the water molecule itself is unfavorable.



(a) Favorably contributing placed water molecule (ΔG$_{HYDE}$ = -1.88 kJ mol^{-1}, MappedΔG$_{HYDE}$ = -5.2 kJ mol^{-1}) with four H-bonds arranged in an almost perfect tetrahedral geometry (PBDid 1h61[284]).

(b) Unfavorably contributing placed water molecule (ΔG$_{HYDE}$ = 1.69 kJ mol^{-1}, MappedΔG$_{HYDE}$ = -2.0 kJ mol^{-1}) due to two low quality H-bonds: one to the backbone nitrogen atom of Asn A 193 and one to the backbone oxygen atoms of Arg A 198 (PBDid 2c78[285]).

Figure 4.13: Example cases for single and mapped HYDE scores of placed water molecules; Left figures = atoms colored by HYDE coloring scheme (green/white/red = favorably/neutral/unfavorably contributing); Right figures = colored by atom colors; Blue mesh = 2fofc electron density grid at 1σ; Yellow lines = H-bonds (hydrogen bonds) formed by the central water molecule.



(c) Favorably contributing placed water molecule (ΔG_{HYDE} = -1.0 kJ mol^{-1}) due to two geometrically good H-bonds to Lys A 761 and Glu A 700 but overall unfavorable surrounding (MappedΔG_{HYDE} = 8.0 kJ mol^{-1}) due to non-optimal H-bonds to the backbone oxygens of His A 679 and Ile A 682 (PBDid 2qoc[286]).

(d) Unfavorably contributing placed water molecule (ΔG_{HYDE} = 1.67 kJ mol^{-1}) and unfavorable surrounding (MappedΔG_{HYDE} = 3.5 kJ mol^{-1}) due to three H-bonds with non-optimal geometries (PBDid 3t6i).

Figure 4.13: Continued.

The energetic contribution of the water molecule itself does not necessarily contain all information needed to evaluate the overall 'happiness', i.e. the overall energetic contribution, of the water molecule. Independently of the water contribution itself, the surrounding can be energetically unfavorable (Figures 4.13c and 4.13d). The first example (Figure 4.13c) shows a water molecule with four hydrogen bonds, two ideal and two non-ideal geometries. Thus, the water molecule itself is energetically favorably contributing, while the backbone oxygens of His and Ile are penalized and lead to an overall unfavorable surrounding. In the second example (Figure 4.13d), the water molecule itself is penalized due to three non-optimal hydrogen bond geometries. The backbone oxygen atom of Thr only participates in the hydrogen bond with the water molecule, while the second hydrogen bond function is covered and cannot interact at all. Therefore, the overall energetic contribution is unfavorable. Similarly, the backbone oxygen of Ala contributes in two hydrogen bonds, both with non-ideal geometries. Thus, the oxygen atom gets an unfavorable energy contribution.

The four examples show that it is important to not only consider the energy contribution of the water molecule itself, but also its direct surrounding. Especially in the drug development process, this aspect can help to identify water positions which lead to an overall gain in affinity upon displacement.

5
Applications and Use Cases

Method development is only one side of the coin. Application scenarios are needed to demonstrate the practical usefulness of the developed methods. Diverse applications were conducted using the different developed methods.

A comparison to state-of-the-art software solution for water placement and scoring based on relevant drug targets shows the advantages as well as disadvantages compared to other methods. Further applications concern the HYDE scoring function.[19,287] On the one hand it was applied according to its original purpose – for virtual screening applications. On the other hand HYDE was used beyond its original scope – for the prediction of amino acid side chain mutations as well as protein-protein interface classification.

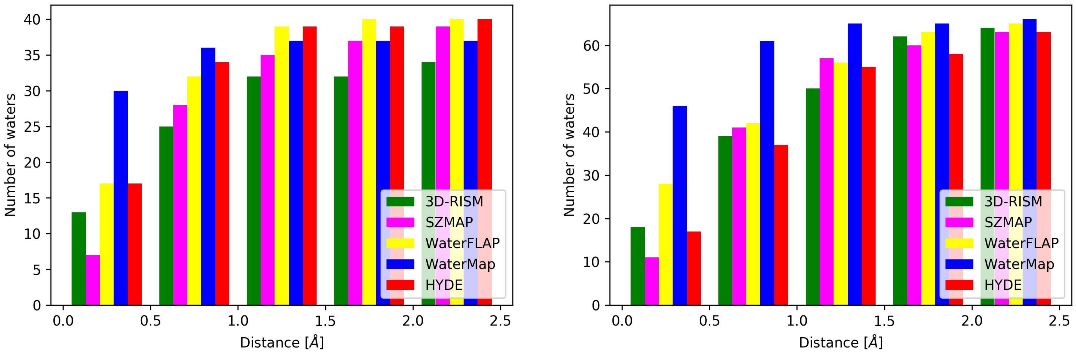
5.1 HYDE$_{\text{water}}$ – Comparison to State-of-the-Art Software Solutions

The developed water placement procedure was compared to state-of-the-art software solutions – 3D-RISM,[162,163] SZMAP,[165] WaterFLAP,[161] and WaterMap.[169,170] Four different measures were introduced as criteria for the comparison of water placement and scoring tools: (1) The distance between crystallographically observed and predicted water molecules; (2) The re-creation of the water$\cdots$water distance network; (3) The total number of placed water molecules in a defined area of interest compared to the number of crystallographically observed water molecules; (4) The correlation between observed SAR and predicted water energies.

Two protein families – Bromodomains (BRD) and Bruton's Tyrosine Kinases (BTK) – with nine crystal structures each were selected for the comparison. Five observed water molecules were analyzed for BRD and eight for BTK structures. The examined ligands affect the water network, which was related to observed changes in SAR. In this section the results obtained with the water placement procedure and HYDE will be discussed. For more details on targets, experiments, and methods see P2.

Distance between Predicted and Crystallographically Observed Water Molecules

The distances between the oxygen atoms of observed and placed water molecules were measured to quantify the accuracy of the placement procedure. WaterMap achieved the highest accuracies for both analyzed protein families within 0.5 Å distance to crystallographically observed water molecules (Figure 5.1). HYDE was the second best for BRD structures with 80% correctly placed water molecules within 1.0 Å distance (Figure 5.1a). WaterFLAP placed water molecules showed the second best accuracies within 0.5 Å to observed water molecules in BTK structures. However, within 1.0 Å distance, all tools (except WaterMap) achieved similar accuracies around 60% (Figure 5.1b).



(a) Distance distribution for placed water molecules in nine BRD structures (total = 41 water molecules).

(b) Distance distribution for placed water molecules in nine BTK structures (total = 66 water molecules).

Figure 5.1: Comparison of water placement procedure with state-of-the-art software solutions. X-axis: distances between crystallographically observed and placed water molecules.

Number of Predicted Water Molecules

For both targets, BRD and BTK, the region of the binding site was approximated with spheres. This way, an 'area-of-interest' was defined in which the number of crystallographically and placed water molecules were counted and compared quantitatively. In both analyzed protein families, HYDE places the same amount of water molecules as crystallographically observable (Table 5.1). Herein, one has to point out that the total number of water molecules placed for each structure individually does not always match exactly. Compared to HYDE, WaterMap also places roughly the same amount of water molecules as observed experimentally. SZMAP and WaterFLAP place 30-60% more water molecules than crystallographically observed in this area, which makes the interpretation of results more difficult. 3D-RISM places less water molecules in BRD and more in BTK structures compared to crystallographically observable.

Table 5.1: Total number of placed water molecules across all nine BRD and BTK structures.

Protein	X-ray	3D-RISM	SZMAP	WaterFLAP	WaterMap	HYDE
BRD	48	42	74	67	43	48
BTK	56	70	72	89	58	56

Re-creation of the Water Network

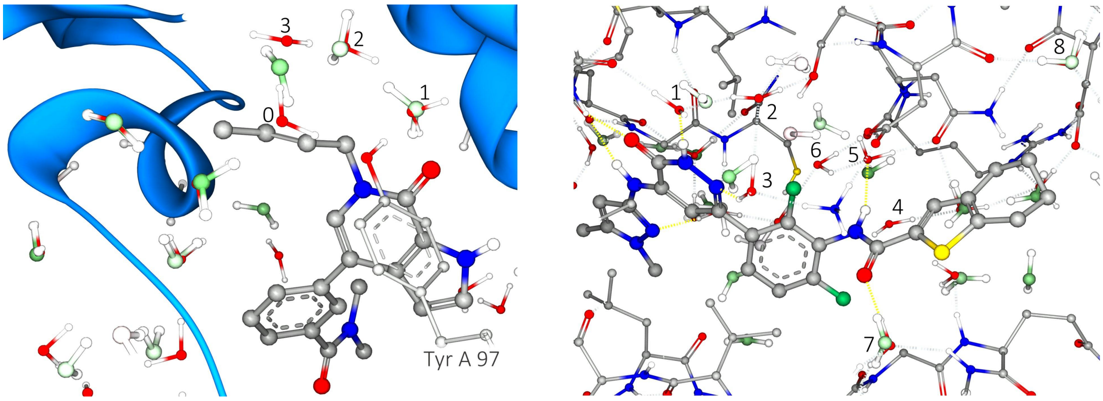
As a basis for the water network re-creation, the pair-wise distances between the oxygen atoms of the crystallographically observed water molecules were calculated ('basic water network'). Subsequently, all pair-wise distances were measured for the different programs and by RMSD measurement compared to the basic water network (Table 5.2). 3D-RISM and SZMAP have larger RSMD values for BRD structures, while SZMAP and WaterFLAP were less accurate in re-creating the water network in BTK structures. WaterMap re-creates the water networks well, except for four BRD structures, where WaterMap was not able to locate one water molecule correctly, which mediates between protein and ligand. HYDE performed well in re-creating the water network in both targets with only two exceptions and RMSD values greater than 1.0 (complex BRD4 compound 4, Figure 5.2a and complex BTK compound 13, Figure 5.2b). In the BRD4 complex, a water molecule was placed in between of two observed ones (water molecules #0 and #3) while the remaining water molecules were placed with high accuracy. In the complex of BTK with compound 13, the water cluster consisting of three crystallographically observed water molecules (#4, #5, and #6) is not predicted well. This is due to the shape of the pocket, which is open to the protein surface in that area. Our water placement procedure relies on interaction directions from protein or ligand atoms and not water alone. Thus, water molecules in this area are less constrained due to less available interaction points.

Table 5.2: Median (average) RMSD values of the re-created water network in BRD and BTK structures.

Protein	3D-RISM	SZMAP	WaterFLAP	WaterMap	HYDE
BRD	1.16 (1.53)	0.88 (0.84)	0.58 (0.61)	0.69 (0.64)	0.48 (0.51)
BTK	0.69 (0.73)	0.88 (0.96)	0.94 (1.04)	0.43 (0.48)	0.69 (0.69)

Energetic Contribution of Water Molecules and SAR Consistency

Based on an overlay of BRD9 structures, the ligand alterations (hydrophobic extensions using carbon chains) do not affect the water molecule network. Thus, the experimentally observed affinity changes should not be due to the water network, but rather the hydrophobic interactions of ligand with



(a) BRD4 compound 2 with crystallographic water molecules and HYDE predicted ones (PDBid 5i88[288]).

(b) BTK compound 13 with crystallographic water molecules and HYDE predicted ones (PDBid 6bln, P2).

Figure 5.2: Comparison of observed (ball-and-stick water molecules with red oxygen atoms) and predicted water molecules (HYDE-colored); Hydrogen bond network was optimized using Protoss.[276]

protein. HYDE predicted individual water molecule contributions are rather favorable for all five water molecules of interest in the active site of BRD9 (Figure 5.3). One exception is BRD9 with compound 1, in which water molecule #1 received a higher energy score ($\Delta G_{HYDE} \approx 0kJmol^{-1}$). Compared to the remaining three structures, in which water #1 is scored favorably by forming three ideal hydrogen bonds, water #1 in BRD9 with compound 1 can only form two geometrically ideal hydrogen bonds. The water molecule was placed in between the two electron pairs of the amide oxygen atom, which is geometrically less favorable. Water #0, which is tetrahedrally coordinated with three hydrogen bonds to protein atoms, is scored more favorable than water #1 across all BRD9 structures. Water #4 mediates between two backbone oxygen atoms and is rotationally restricted. Therefore, water #4 is scored most consistent between the different BRD9 structures.

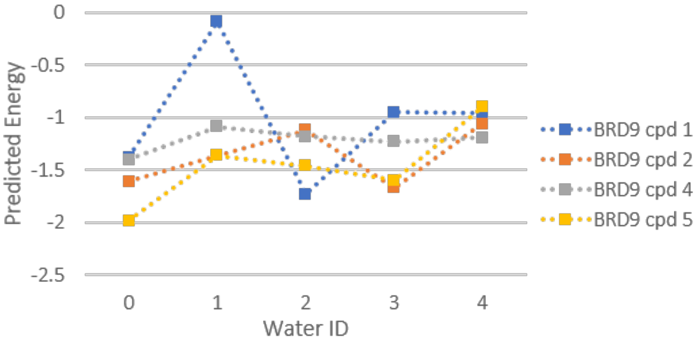
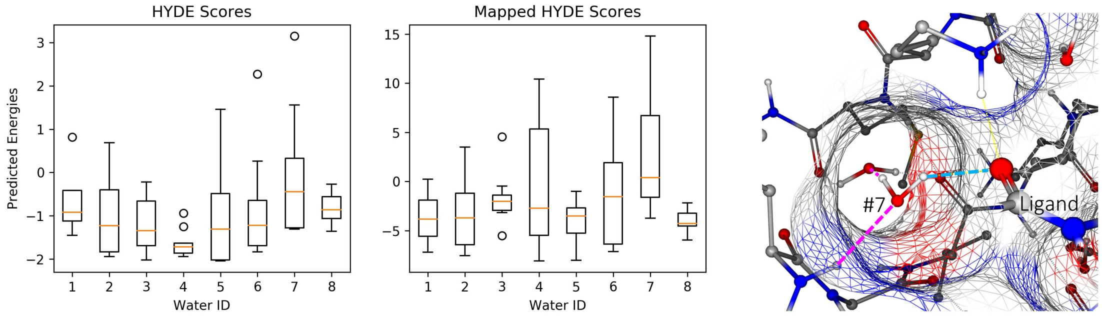


Figure 5.3: HYDE predicted energies for placed water molecules in BRD9 structures with different compounds (cpd); Connecting dashed lines are for visualization purpose only.

Different aspects were analyzed for the water energies in BTK structures. Experimentally, the displacement of water #1 resulted in a decrease in affinity. Thus, water #1 was expected to contribute favorably to the overall binding affinity. This aspect is well represented by the HYDE score for water #1 (Figure 5.4a). Also the surrounding of water #1 is predicted to be favorable (Figure 5.4a *mapped HYDE scores*). Water #8 is well integrated into the protein. It was unsuccessfully tried experimentally to replace this water molecule. The HYDE score for the water itself as well as the HYDE scores of the surrounding are favorable. Water #7 is present in all structures. Compared to water #8, it is spatially more restricted. More than half of its pocket is apolar and the remaining interaction partners are a water molecule, a backbone nitrogen and an amide oxygen atom of the ligand (Figure 5.4b). However, both the backbone and ligand atoms are not arranged ideally. Therefore, water #7 has a greater variance in its predicted binding affinity. Water molecules #2 and #3 are close to water #1. Thus, their scores vary more due to the displacement of water #1. Waters #4, #5, and #6 form one cluster which is located at the rim of the pocket, which leads to a greater distribution of their scores.



(a) Variability of HYDE predicted water energy scores across eight water molecules for nine BTK structures; Box: line = median, boundaries = upper and lower quartile; Whiskers = minimum and maximum HYDE score for the individual water molecules; Circles = Outliers.

(b) Water #7 of BTK compound 9; Mesh = active site surface colored by atom type; Magenta/Blue dashed line = ideal/non-ideal H-bond.

Overall, the developed water placement procedure proved to be comparable or better than three out of four state-of-the-art commercial software solutions. WaterMap is superior in placing water molecules in close distance to the crystallographically observed ones. The placement of water molecules works fairly well for all analyzed software solutions (57-86% for BRD and 60% for BTK within 1.0 Å distance to observed water molecules).

However, the predicted water energies by 3D-RISM, SZMAP, WaterFLAP, and WaterMap did not correlate with experimental SAR. Inter- as well as intra-tool results were inconsistent. Especially WaterMap scored highly integrated water molecules unfavorable, which was not as expected. HYDE predicted water energies have a high dependency on the hydrogen bond network and by this on the number of formed hydrogen bonds. Even though the predicted water energies were not as consistent for BRD as suggested by the experimental structures, the HYDE scores for BTK structures could explain some of the hypothesis derived from experimentally measured binding affinities. For more details on the results of the state-of-the-art software solutions please refer to P2.

5.2 HyPPI – Protein-Protein Interface Classification

Protein-protein interactions (PPI) display an interesting target for drug development because they are often involved in signal transduction pathways.[289] PPIs can be differentiated according to their stability into obligate and non-obligate PPIs. The separate protein units of obligate PPIs are not stable on their own, whereas non-obligate protein units are stable in their unbound form. Further, PPIs can be classified based on their lifetime into permanent and transient interactions. Permanent PPIs are represented by complexes only stable in their interacting form while transient ones interact when their function is needed.[290] In order to inhibit a protein-protein interaction it is necessary to differentiate the interface into permanent or transient, with the latter displaying an ideal target for drug design. The available data of PPIs resulting from X-ray crystallization poses another challenge: The observed PPIs need to be discriminated into true biological interfaces and into artificial contacts due to the crystallization process. Crystal contacts only exist due to the protein crystallization process and do not possess any biological function.

Different solutions, based on free energy calculations or classification models using geometric and physico-chemical descriptors, have been developed for the discrimination of PPIs.[291–301] Many methods distinguishing between crystal artifacts and biological complexes achieve high accuracies, generally above 77%. Also methods for the classification of non-obligate and obligate PPIs or permanent and transient PPIs reach accuracies of 92%. However, large amounts of descriptors, ranging from 7 up to 213, were used for the differentiation.

Our classification of PPIs – HyPPI[c] – is based on two descriptors only: (1) the hydrophobicity of the interface based on the HYDE hydrophobic dehydration score (Equation 4.8) and (2) the accessible surface area difference between the unbound and bound state (Equation 5.2).

$$IFR(x) = \frac{MS_{IF}(x)}{MS_{unbound}(x)} \tag{5.1}$$

$$IF_{quotient} = \frac{min(IFR(A), IFR(B))}{max(IFR(A), IFR(B))} \tag{5.2}$$

The interface ratio of protein x (IFR(x)) is calculated using the molecular surface of the protein interface (MS$_{IF}(x)$) and the whole surface area of the unbound protein (MS$_{unbound}(x)$). The interface quotient (IF$_{quotient}$) is formed by the smaller IFR of the PPI divided by the larger IFR (Figure 5.5). Those

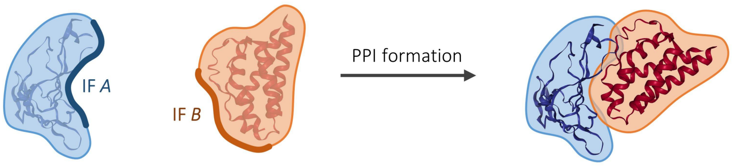


Figure 5.5: Exemplary formation of a PPI; Bold lines = change in accessible surface area upon PPI formation.

[c] Developed during my bachelor thesis: Vennmann, E. Klassifikation von Protein-Protein-Komplexen auf Basis der Bewertungsfunktion HYDE. B.Sc. Thesis, Universität Hamburg, 2010.

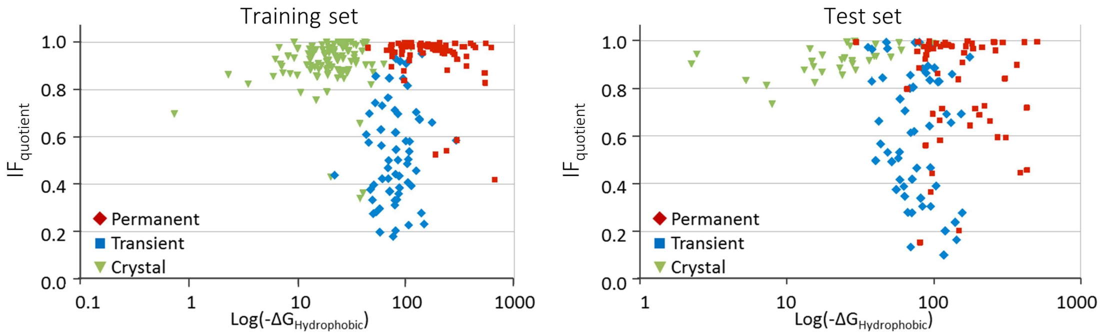
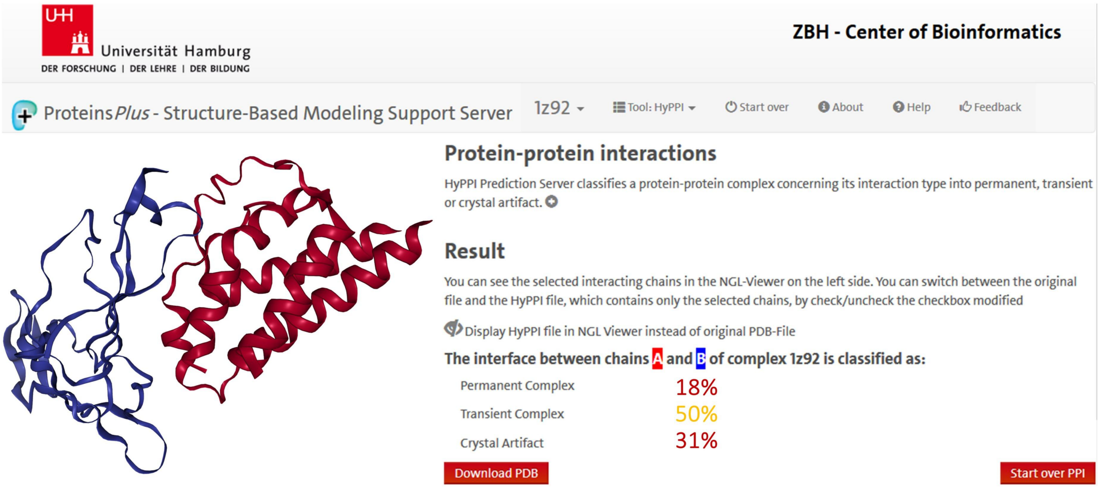


Figure 5.6: Distribution of descriptors used for SVM-based classification of PPI for the training (number of complexes: C = 120, P = 74, T = 60) and test set (number of complexes: C = 32, P = 59, T = 61).

two descriptors were used in a two-stage support vector machine (SVM, R package e1071[302]). In a first discrimination step, crystal artifacts were separated from biological relevant PPIs. The following step classified biological PPIs into transient and permanent ones. Thus, with a rather simple set of descriptors a sufficient discrimination of PPIs into crystal artifacts, permanent, and transient, could be achieved (92.5% on a training set and 77.9% on an independent test set, Figure 5.6). An example for the transient PPI of interleukin-2 and its receptor (IL-2/IL-2Rα) is given in Figure 5.7. It has been shown that the PPI of IL-2/IL-2Rα can be inhibited with small molecules.[303] Thus, it is correctly classified as transient. For more information see D6 and D7.

At the time when HyPPI was developed, no adequate handling of water molecules in HYDE was available. Therefore, it would be interesting to test the discrimination of PPIs based on the latest development of HYDE, including the water placement procedure.


UHH
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG
ZBH - Center of Bioinformatics
ProteinsPlus - Structure-Based Modeling Support Server
1z92
Tool: HyPPI
Start over
About
Help
Feedback
Download PDB
Start over PPI

Protein-protein interactions

HyPPI Prediction Server classifies a protein-protein complex concerning its interaction type into permanent, transient or crystal artifact.

Result

You can see the selected interacting chains in the NGL-Viewer on the left side. You can switch between the original file and the HyPPI file, which contains only the selected chains, by check/uncheck the checkbox modified

Display HyPPI file in NGL Viewer instead of original PDB-File

The interface between chains A and B of complex 1z92 is classified as:

Permanent Complex	18%
Transient Complex	50%
Crystal Artifact	31%

Figure 5.7: Screenshot of the Proteins*Plus* Server[84] interface of HyPPI; Example structure: interleukin-2 and its receptor (IL-2/IL-2Rα, PDBid: 1z92[304]).

5.3 HYDE$_{\text{protein}}$ – Amino Acid Mutation Predictions

Optimization of enzyme functionality is of great interest in biotechnological processes. Oftentimes amino acid side chain mutations are exploited to increase enzyme stability, turnover and yield rate, or alter its substrate specificity.[305]

The prediction of energetic effects of amino acid side chain mutations using the HYDE scoring function can be performed with HYDE$_{protein}$. Due to the generic physics-based concept of HYDE and no training on experimental binding affinity data, it can be applied to score mutation effects without alteration to the underlying scoring function.

HYDE$_{protein}$ applies an enumeration approach to sample possible conformations of the mutated amino acid (Figure 5.8). Those conformations are considered as staring points for GeoHYDE optimization. Since GeoHYDE is a local optimization, different starting points need to be evaluated. The optimized amino acid side chain are then scored with HYDE$_{protein}$ and the best scored conformation is selected. Different experiments were conducted to test HYDE$_{protein}$: (1) remutation experiments, (2) cross-mutation experiments, and (3) protein stability predictions.

Remutation experiments are commonly used and allowed a comparison to state-of-the-art methods. HYDE$_{protein}$ showed comparable accuracies to other studies based on remutation and protein stability prediction experiments. The cross-mutation experiments were newly introduced to examine more realistic scenarios, i.e. a mutation of one amino acid into another one. As expected, those experiments led to a decrease in accuracy, especially if multiple mutations that effected each other occurred. For more details on the mutation procedure and comparisons to state-of-the-art methods see D8.

At the time when HYDE$_{protein}$ was evaluated, the handling of water molecules in HYDE was limited. Either crystallographically observed water molecules were included or excluded. However, especially for mutation studies with varying sizes of amino acid side chains, water molecules are necessary. Herein, the mutation of a larger to a smaller amino acid might create enough space for a water molecule, while the other way round a water molecule might be displaced. Therefore, it would be interesting to analyze the impact of water molecules on the accuracy of side chain conformation predictions as well as their energetic contributions.

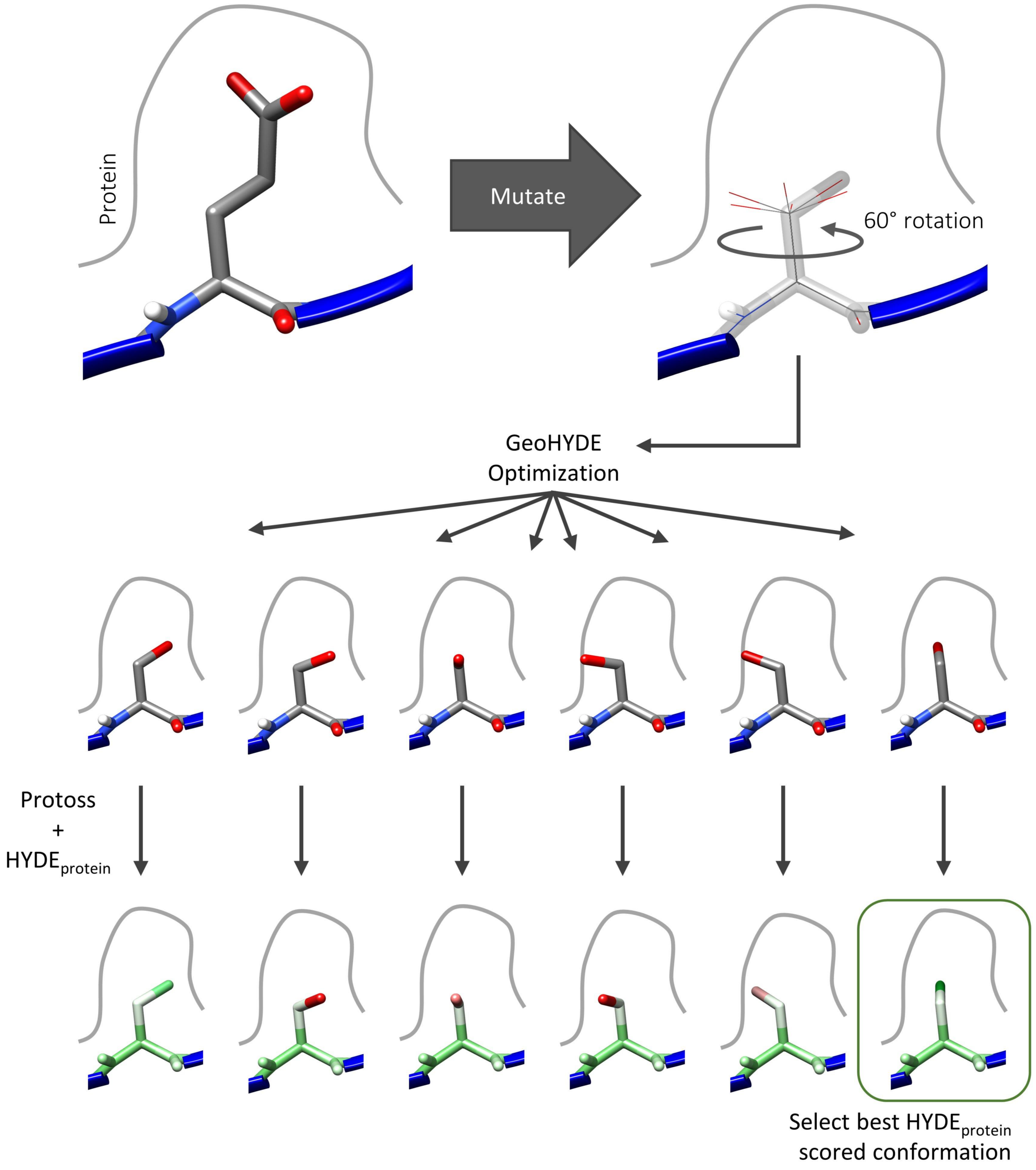


Figure 5.8: Overview of amino acid side chain mutation process; Protoss: Hydrogen bond network optimization.[276]

5.4 Biotechnological Applications

HYDE has been used in diverse scenarios relevant for biotechnology, from identification of a natural substrate to analysis and alteration of substrate specificity.

5.4.1 EstN2 – Elucidation of Enzyme Functionality

The enzyme EstN2 (Figure 5.9) was discovered by genome analysis of the archaeon *Candidatus* Nitrososphaera gargensis, which was obtained from terrestrial hot springs.[306] However, the actual function of EstN2 remained unclear.

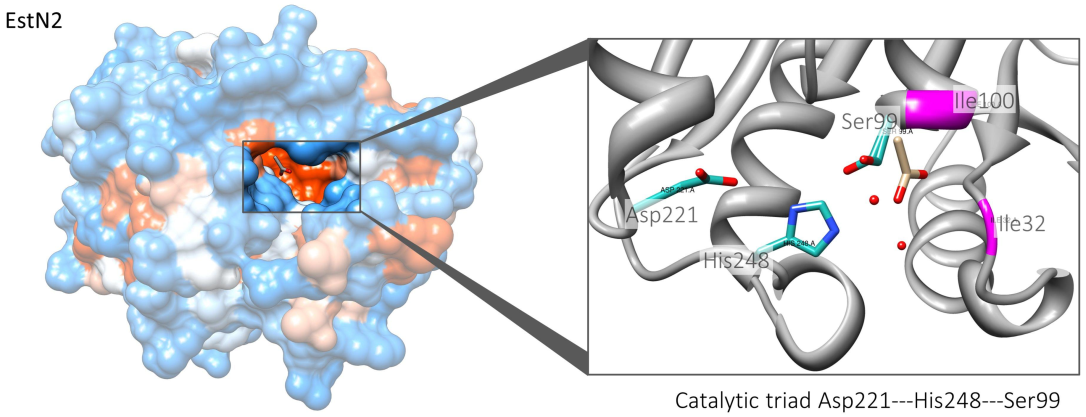


Figure 5.9: Active site of EstN2 with its catalytic triad (Molecular graphics were created using UCSF Chimera[65]).

Structural alignment revealed chloroperoxidase, enol-lactonase, and esterase as closest homologous proteins. Thus a virtual screening of potential substrates from other known enzymes was conducted. In a first attempt a data set of 3036 ester and lactone containing small molecules was compiled based on BRENDA (BRaunschweig ENzyme DAtabase).[307] This dataset was then docked with FlexX[208] and scored using HYDE (Figure 5.10).

The highest ranked molecules often contained glycerol derivatives (Figure 5.10: Results 1). Therefore, it was concluded that a potential function of EstN2 was the degradation of phosphoglycerols. Within the top 100 scored molecules enole containing molecules were only represented once. Thus, enoles were unlikely to be a natural substrate of the enzyme. This finding was supported by laboratory experiments that could not detect any enzyme activity using enoles as substrates.

In a second approach suggested small molecules were examined for their fit in the active site of EstN2 (Figure 5.10: 2. Screening). However, the molecules showed rather fragment like characteristics, i.e. molecular weight 130-152 Da, and only occupied a fraction of the binding site (pocket size 585 Å^3, molecule sizes: 126-192 Å^3). Thus, it was difficult to achieve definite conclusions from this second approach because multiple binding modes were scored similarly (Figure 5.10: Results 2).

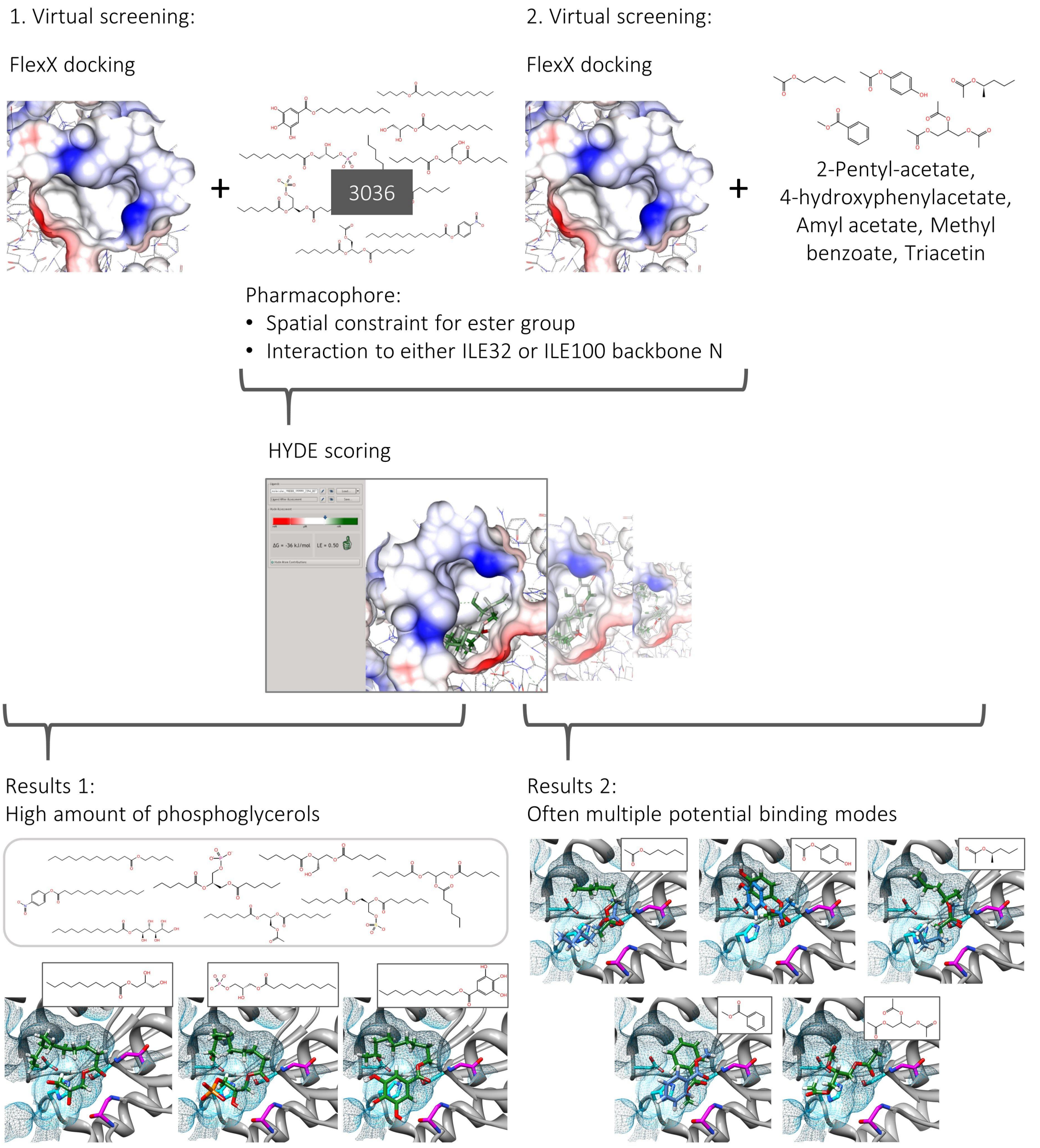

1. Virtual screening:
FlexX docking
+
3036
2. Virtual screening:
FlexX docking
+
2-Pentyl-acetate, 4-hydroxyphenylacetate, Amyl acetate, Methyl benzoate, Triacetin
Pharmacophore:
• Spatial constraint for ester group
• Interaction to either ILE32 or ILE100 backbone N
HYDE scoring
Ligands
Ligand After Assessment
Load...
Save...
Hyde Assessment
ΔG = -36 kJ/mol
LE = 0.50
Hyde Atom Contributions
Results 1:
High amount of phosphoglycerols
Results 2:
Often multiple potential binding modes

Figure 5.10: Virtual screening processes for the elucidation of the natural substrate of EstN2.

5.4.2 Lipase Cal B – Analysis of Substrate Specificity

Lipase Cal B catalyzes the reaction of methyl glucoside with lauric acid to methyl glucoside laurate (Figure 5.11). Lipase Cal B exhibits a high substrate specificity. The removal of the methyl group, i.e. from methyl glucoside to α-D-glucose, leads to a complete loss of enzyme activity (Figure 5.12 top). Thus a virtual screening approach was used to generate explanations for this observation.

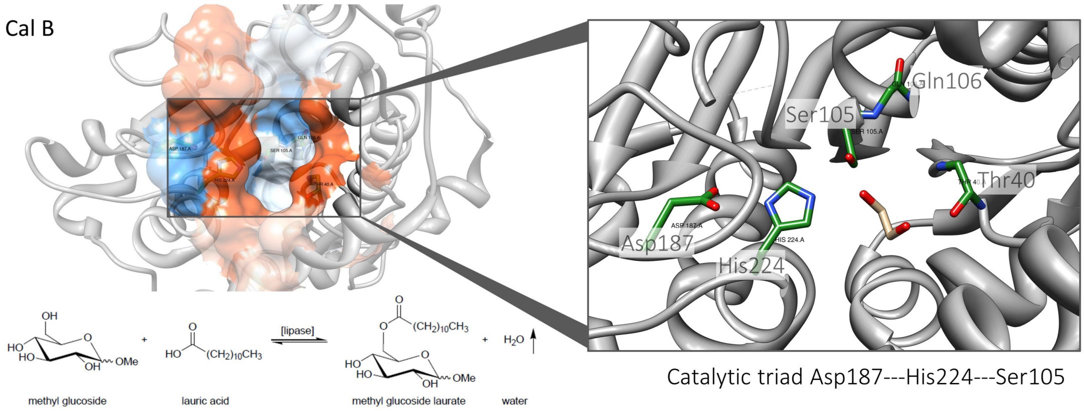


Figure 5.11: Active site of lipase Cal B with its catalytic triad (Molecular graphics were created using UCSF Chimera[65]).

In a first approach the final product, methyl glucoside laurate or α-D-glucose laurate, was docked using different spatial constraints. However, none of these approaches let to satisfying results. Thus, a different approach was pursued, using a step-wise approach (Figure 5.12 middle). The first docking consisted of placing lauric acid into the active site. Since lauric acid has a long and flexible carbon chain, its length was shortened. Using the best HYDE scored conformation of the fatty acid, either methyl glucoside or α-D-glucose were docked into the active side including the pre-docked fatty acid. In both docking steps a pharmacophore was used to enforce the contact of catalytic side chains with the functional groups from the educts.

Overall, the substrate specificity might be explained by a hydrophobic effect of the methyl group from the methyl-glucoside pointing towards the carbon-chain of lauric acid. The energetic advantage for methyl glucoside of $\Delta\Delta G_{HYDE}$ -3 kJ mol^{-1} and -5 kJ mol^{-1} for hexanoic and decanoic acid, respectively, might even further increase for lauric acid (Figure 5.12 bottom). The hydrophilic hydroxyl group of α-D-glucose on the other hand would point towards the aliphatic chain and would lead to a repulsion.

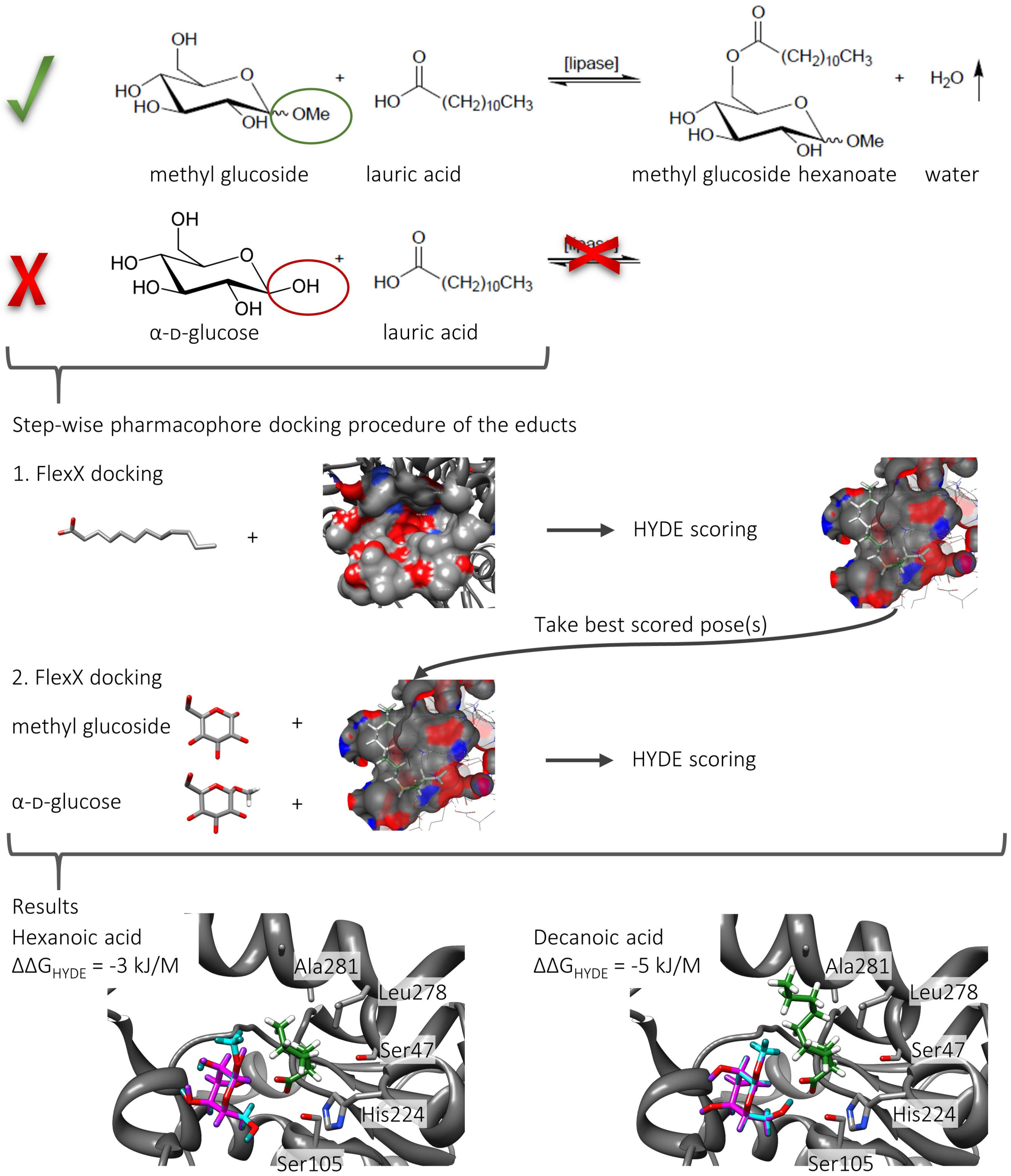


Figure 5.12: Docking and scoring processes of Lipase Cal B; based on X-ray structure 4k6g (Molecular graphics of the results section of the figure were created using UCSF Chimera[65]).

5.4.3 Trimethlyguanosine Synthase – Substrate Specificity Alterations

Trimethlyguanosine synthases (TGSs) are responsible for methylation of the 5′-Cap of mRNA.[308] Herein, a methyl group of the cofactor S-adenosylmethionine is transferred to the 5′-Cap of mRNA. TGSs are an interesting target for further analysis of the underlying mechanisms and eventual influence of mRNAs. Therefore, analogs of the cofactor can be used to transfer other functional groups than methionine to allow chemical click-reactions. Those reactions can subsequently be used to label the mRNA and visualize them, i.e. using fluorescence labeling.

The aim of this analysis was the alteration of substrate specificity of TGS of the organism *Gardia lamblia*. First, HYDE was used to predict the relative affinities of the cofactor in its methylized (SAM) and demethylized (SAH) form, as well as its analog AdoPropen (Table 5.3 and Figure 5.13a). HYDE was able to score the different small molecules correctly, i.e. their relative scores in human and *G. lamblia* TGS, hTGS and glaTGS respectively, were as observed by experiment (compare Table 5.3 and Figure 5.13b).

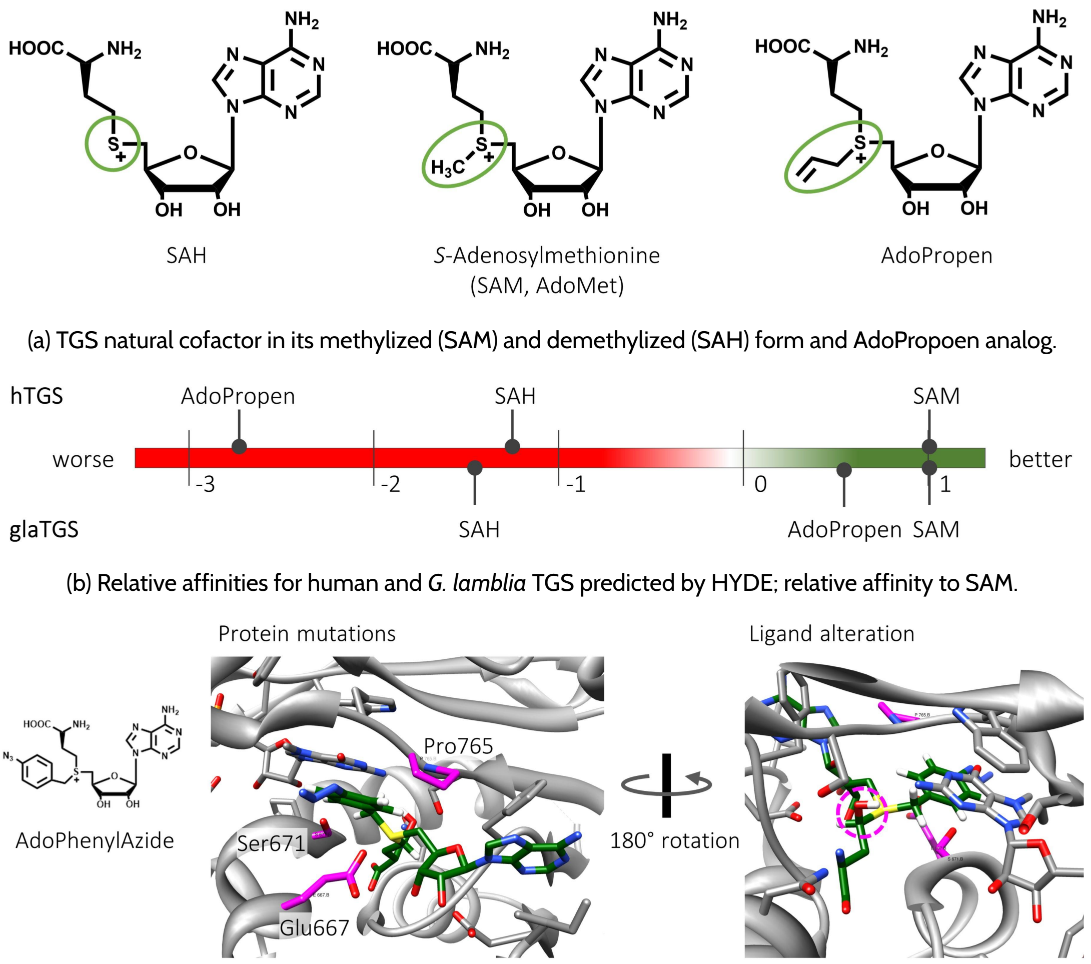
Table 5.3: Substrate and analoga activity of TGS; natural cofactor S-adenosylmethionine (SAM) and its demethylized form (SAH) compared to an analog with propen (AdoPropen); SAH as reference point; hTGS = human protein; glaTGS = *G. lamblia* form of TGS.

Protein	SAH	SAM	AdoPropen
hTGS	0	+++	-
glaTGS	0	+++	+

Next, the SAM binding site was examined to suggest amino acid mutations to increase the binding site to accept bigger substituents than methyl, i.e. propen or even phenyl azide. Since the protein structure of glaTGS was based on a homology model of hTGS, the reliability of suggested amino acid mutations for the X-ray hTGS form was higher (Figure 5.13c and Table 5.4). Especially for the stabilization of the mostly hydrophobic analogs, smaller but hydrophobic amino acids such as alanine or glycine were suggested. Ligand binding affinity could be enhanced by integrating suitable hydroxyl groups for further hydrogen bonding.

Based on glaTGS our collaboration partner could achieve some success with directed alterations of mRNA. However, due to specific pre-methylation conditions needed by glaTGS, a different enzyme was finally chosen, Ecm1, which provided a greater promiscuity for cofactors and their analogs.[309]

At the time, when this study was conducted, neither HYDE$_{protein}$ nor the water placement and scoring procedure was available. Both developments would now allow a more detailed examination of amino acid alterations.



(a) TGS natural cofactor in its methylized (SAM) and demethylized (SAH) form and AdoPropoen analog.

(b) Relative affinities for human and *G. lamblia* TGS predicted by HYDE; relative affinity to SAM.

(c) Proposed amino acid mutations (magenta amino acids) and ligand alteration (magenta dashed circle) for selectivity advantages (structure: hTGS with AdoPhenylAzide; Molecular graphics were created using UCSF Chimera[65]).

Figure 5.13: Analysis of human and *G. lamblia* TGS.

Table 5.4: Proposed amino acid mutations and ligand alteration for hTGS according to Figure 5.13c.

Alteration		Reason
From	To	
Ser671	Ala or Gly	steric advantages
Glu667	Ile or Val	stabilization of phenyl ring
Pro765	Ala or Gly	steric advantages
H	OH	hydrogen bond between ligand and protein

6
Conclusions and Further Directions

In the presented work a reliable procedure for placing water molecules in protein structures was developed. First, an automatic assessment of protein structures was evolved – EDIA and EDIA$_m$. Second, *NAOMI*nova was implemented for the deduction of interaction directions and geometries. The derived interaction geometries were exploited to detect suitable available space in protein structures for water molecules. Subsequently, the identification of free space was utilized for placing explicit water molecules. Finally, those predicted water positions were scored with HYDE to predict their energetic contributions.

Great emphasis was put on the use of as much experimental data as possible. Herein, structural data as well as their underlying electron density data was exploited. The automatic structure quality assessment – EDIA and EDIA$_m$ – provide an objective criterion for the evaluation of protein structures without the subjective, visual interpretation of electron density grids. In addition to the qualitative assessment, the EDIA was applied to compile a large data set of well resolved water molecules for the evaluation of the developed water placement strategy. Based on protein crystal structures, a large-scale analysis of interaction directions was conducted. With the developed tool *NAOMI*nova 22 functional groups, present in amino acids as well as ligands, were evaluated for their preferred interaction directions. The analysis showed that donor interactions agree well with theoretical considerations, while acceptor functions can vary significantly. Furthermore, unexpected geometries could often be ascribed to structural artifacts, such as close atom contacts around metals. The derived interaction directions and geometries were finally used for several aspects: (1) the identification of free space, (2) subsequently for the water placement procedure, and (3) for scoring of hydrogen bond interactions in HYDE. The developed identification of free space accurately represents areas in protein structures, where the full volume of a water molecules would fit. Polar atoms that can form hydrogen bonds to a potential water molecule were included with smaller atom radii. The reduction of the radii was derived from a water$\cdots$atom distance analysis with *NAOMI*nova. Apolar atoms on the other hand did not show closer distances than expected by their van-der-Waals radii and were treated accordingly. The placement of explicit water molecules from discrete points used for the identification of free space was done with the development of a self-assembly procedure. The potential water positions defined by the interaction surfaces cluster themselves and thus result in explicit water positions. These water positions were finally scored with HYDE. Herein, diverse aspects of the HYDE scoring function were adapted to handle water molecules more accurately.

Several aspects influence the developed water placement procedure. Initially, Protoss[276] is used for the optimization of the hydrogen bond network and thus defines the starting point for the identification of free space. If the hydrogen bond network is optimized differently, the starting positions could result in differently placed water molecules. In addition to this, alternate conformations assigned for amino acids or ligands in protein structures lead to a different hydrogen bond network, different interaction directions and thus, different locations for the predicted water networks. However, protein structures only represent a snapshot of the flexibility of the protein structure in nature. Especially water molecules and their networks are highly variable. This aspect could potentially be used to analyze the effects of protein flexibility on the water network and its corresponding energetic contribution.

The integration of the developed identification of free space (FSI) and water placement (WarPP) have different consequences – positive as well as negative. The estimation of the dehydration probability (p_{dehyd}) based on the FSI now allows a detailed sampling of the interaction surface in ideal hydrogen bond distance as well as an accurate representation of a water molecule. However, a complete representation would demand sampling of the full hydrogen bond distance range. This would eliminate the discrete step in ideal angle range in the scoring function but would greatly increase the run-time, especially for the optimization with GeoHYDE.

Differences in modeled protein structures can lead to a correct water placement in one and none in another structure due to steric hindrances. Thus, the predicted HYDE scores can differ. In the first case, hydrogen bonds can be formed and the participating atoms are scored favorably, in case of good hydrogen bond qualities. The latter one has unsatisfied hydrogen bond functions, which are penalized by HYDE. An increase in scoring robustness could be achieved by placing water molecules after an initial protein-ligand minimization based on the HYDE scoring function. Tight areas within the binding site might be opened to accommodate water molecules that were not available pre-optimization. A second geometric optimization would be run after the water placement to ensure high-quality hydrogen bond geometries. This adaption would increase the run-time and might not be necessary or beneficial for all protein-ligand structures. It would only be useful if many unsatisfied hydrogen bond functions are within the active site. Those might be satisfied by forming hydrogen bonds to a water molecule, which does not have enough space without an initial optimization step.

Theoretical considerations reason the free energy contribution of water molecules to be between -0.7 and -2.2 kcal mol^{-1} (-2.9 to -9.2 kJ mol^{-1}).[70,71] The maximum favorable energy contribution for a water molecule predicted by HYDE is -2.05 kJ mol^{-1}. This number is derived from the theoretical concept behind the HYDE scoring function and is at the lower range of the aforementioned theoretical considerations.

The diverse application scenarios showed the practicality of the water placement procedure as well as HYDE itself. Especially the HYDE studies, i.e. $\text{HYDE}_{protein}$, HyPPI, and the biotechnology examples, were conducted when the development of HYDE_{water} was not finished. A comparison to state-of-the-art, commercial software solutions proved the developed water placement and scoring as equally good and in some cases better. Thus, it would be interesting to re-evaluate the HYDE studies including HYDE_{water}. Especially for the mutation prediction, water molecules play an important role, i.e. they might be displaced if a small amino acid is replaced by a larger one or they might mediate interactions, when a larger amino acid is replaced by a smaller one.

The water placement procedure could further be applied to analyze the impact of protein flexibility on the water network and *vice versa*, to consistently place water molecules in docking poses, and to ensure a homogeneous representation of water molecules in protein-ligand structures for scoring. As a next step, the impact of WarPP on protein-ligand, protein-protein, as well as intra protein HYDE scores demands an extensive evaluation. Overall, the developed water placement and scoring procedure can aid the identification of water molecules within a protein-ligand binding site. Water molecules that are good to incorporate during drug design, because they can mediate between protein and ligand and are energetically favorable, and those that are good to target for displacement, because they are weakly integrated and contribute unfavorably, can be differentiated.

Bibliography

(*1*) Cooper, G. M., In *cell - a Mol. approach*, 2nd editio; ASM Press: 2000, p 689.

(*2*) Janin, J., (1999). Wet and dry interfaces: The role of solvent in protein-protein and protein-DNA recognition. *Structure 7*, R277–R279.

(*3*) Langhorst, U., Backmann, J., Loris, R., and Steyaert, J., (2000). Analysis of a water mediated protein-protein interactions within RNase T1. *Biochemistry 39*, 6586–6593.

(*4*) Rodier, F., Bahadur, R. P., Chakrabarti, P., and Janin, J., (2005). Hydration of protein-protein interfaces. *Proteins Struct. Funct. Genet. 60*, 36–45.

(*5*) Ahmad, M., Gu, W., Geyer, T., and Helms, V., (2011). Adhesive water networks facilitate binding of protein interfaces. *Nat. Commun. 2*, 261.

(*6*) Rühlmann, A., Kukla, D., Schwager, P., Bartels, K., and Huber, R., (1973). Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. Crystal structure determination and stereochemistry of the contact region. *J. Mol. Biol. 77*, 417–436.

(*7*) Kawasaki, K., Kondo, H., Suzuki, M., Ohgiya, S., and Tsuda, S., (2002). Alternate conformations observed in catalytic serine of Bacillus subtilis lipase determined at 1.3 A resolution. *Acta Crystallogr. Sect. D Biol. Crystallogr. 58*, 1168–1174.

(*8*) Phillips, R. S., (2002). How does active site water affect enzymatic stereorecognition? *J. Mol. Catal. B Enzym. 19-20*, 103–107.

(*9*) Higgins, M. K., and Lea, S. M., (2017). On the state of crystallography at the dawn of the electron microscopy revolution. *Curr. Opin. Struct. Biol. 46*, 95–101.

(*10*) Blow, D., In *Outl. Crystallogr. Biol.* OUP Oxford: 2002, p 196.

(*11*) Teeter, M., (1991). Water-Protein Interactions: Theory And Experiment. *Annu. Rev. Biophys. Biomol. Struct. 20*, 577–600.

(*12*) Carugo, O., (2016). When proteins are completely hydrated in crystals. *Int. J. Biol. Macromol. 89*, 137–143.

(*13*) Cheng, T., Li, X., Li, Y., Liu, Z., and Wang, R., (2009). Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model. 49*, 1079–1093.

(*14*) Søndergaard, C. R., Garrett, A. E., Carstensen, T., Pollastri, G., and Nielsen, J. E., (2009). Structural artifacts in protein-ligand X-ray structures: Implications for the development of docking scoring functions. *J. Med. Chem. 52*, 5673–5684.

(*15*) Reulecke, I., Adaptives Scoring im strukturbasierten Wirkstoentwurf. HYDE -Eine neue Scoring-Funktion für die konsistente Bewertung von Protein-Ligand Komplexen., Ph.D. Thesis, Universität Hamburg, 2008.

(*16*) Reulecke, I., Lange, G., Albrecht, J., Klein, R., and Rarey, M., (2008). Towards an integrated description of hydrogen bonding and dehydration: Decreasing false positives in virtual screening with the HYDE scoring function. *ChemMedChem 3*, 885–897.

(*17*) Schneider, N., HYDE: Konsistente Bewertung von Protein-Ligand-Komplexen auf der Basis von Wasserstoffbrücken- und Dehydratationsenergie. Von der Theorie zur Anwendung., Ph.D. Thesis, Universität Hamburg, 2012.

(*18*) Schneider, N., Klein, R., Lange, G., and Rarey, M., (2012). Nearly no scoring function without a hansch-analysis. *Mol. Inform. 31*, 503–507.

(*19*) Schneider, N., Lange, G., Hindle, S., Klein, R., and Rarey, M., (2013). A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: Methods behind the HYDE scoring function. *J. Comput. Aided. Mol. Des. 27*, 15–29.

(*20*) Chen, J. M., Xu, S. L., Wawrzak, Z., Basarab, G. S., and Jordan, D. B., (1998). Structure-based design of potent inhibitors of scytalone dehydratase: Displacement of a water molecule from the active site. *Biochemistry 37*, 17735–17744.

(*21*) Wissner, A., et al. (2000). 4-Anilino-6,7-Dialkoxyquinoline-3-Carbonitrile Inhibitors of Epidermal Growth Factor Receptor Kinase and Their Bioisosteric Relationship To the 4-Anilino-6,7-Dialkoxyquinazoline Inhibitors. *J. Med. Chem. 43*, 3244–3256.

(*22*) Seo, J., Igarashi, J., Li, H., Martásek, P., Roman, L. J., Poulos, T. L., and Silverman, R. B., (2007). Structure-based design and synthesis of Nω-nitro-L-arginine- containing peptidomimetics as selective inhibitors of neuronal nitric oxide synthase. Displacement of the heme structural water. *J. Med. Chem. 50*, 2089–2099.

(*23*) Berman, H. M., (2000). The Protein Data Bank. *Nucleic Acids Res. 28*, 235–242.

(*24*) Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., and Adams, P. D., (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. Sect. D Biol. Crystallogr. 68*, 352–367.

(*25*) Touw, W. G., and Vriend, G., (2014). BDB: Databank of PDB files with consistent B-factors. *Protein Eng. Des. Sel. 27*, 457–462.

(*26*) Trueblood, K. N., Bürgi, H. B., Burzlaff, H., Dunitz, J. D., Gramaccioli, C. M., Schulz, H. H., Shmueli, U., and Abrahams, S. C., (1996). Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallogr. Sect. A 52*, 770–781.

(*27*) Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjeldgaard, M., (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. Sect. A 47*, 110–119.

(*28*) Read, R. J., et al. (2011). A new generation of crystallographic validation tools for the Protein Data Bank. *Structure 19*, 1395–1412.

(*29*) Jones, T. A., and Kjeldgaard, M., (1997). Electron-density map interpretation. *Methods Enzymol. 277*, 173–208.

(*30*) Tickle, I. J., (2012). Statistical quality indicators for electron-density maps. *Acta Crystallogr. Sect. D Biol. Crystallogr. 68*, 454–467.

(*31*) Otting, G., Liepinsh, E., and Wuthrich, K., (1991). Protein hydration in aqueous solution. *Science (80-.). 254*, 974–980.

(*32*) Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P. T., Holton, J. M., Echols, N., and Alber, T., (2011). Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc. Natl. Acad. Sci. 108*, 16247–16252.

(*33*) Kaieda, S., and Halle, B., (2013). Internal water and microsecond dynamics in myoglobin. *J. Phys. Chem. B 117*, 14676–14687.

(*34*) Drew, H. R., and Dickerson, R. E., (1981). Structure of a B-DNA dodecamer. III. Geometry of hydration. *J. Mol. Biol. 151*, 535–556.

(*35*) Mattea, C., Qvist, J., and Halle, B., (2008). Dynamics at the protein-water interface from 17O spin relaxation in deeply supercooled solutions. *Biophys. J. 95*, 2951–63.

(*36*) Nakasako, M., (2004). Water-protein interactions from high-resolution protein crystallography. *Philos. Trans. R. Soc. B Biol. Sci. 359*, 1191–1206.

(*37*) Carugo, O., (2017). Protein hydration: Investigation of globular protein crystal structures. *Int. J. Biol. Macromol. 99*, 160–165.

(*38*) Halle, B., and Denisov, V. P., (2001). Magnetic relaxation dispersion studies of biomolecular solutions. *Methods Enzymol. 338*, 178–201.

(*39*) Mattos, C., (2002). Protein-water interactions in a dynamic world. *Trends Biochem. Sci. 27*, 203–208.

(*40*) Combet, S., and Zanotti, J.-M., (2012). Further evidence that interfacial water is the main "driving force" of protein dynamics: a neutron scattering study on perdeuterated C-phycocyanin. *Phys. Chem. Chem. Phys. 14*, 4927.

(*41*) Levy, Y., and Onuchic, J. N., (2006). Water Mediation in Protein Folding and Molecular Recognition. *Annu. Rev. Biophys. Biomol. Struct. 35*, 389–415.

(*42*) Kinoshita, M., (2009). Importance of translational entropy of water in biological self-assembly processes like protein folding. *Int. J. Mol. Sci. 10*, 1064–1080.

(*43*) Mallamace, F., Corsaro, C., Mallamace, D., Vasi, S., Vasi, C., Baglioni, P., Buldyrev, S. V., Chen, S.-H., and Stanley, H. E., (2016). Energy landscape in protein folding and unfolding. *Proc. Natl. Acad. Sci. 113*, 3159–3163.

(*44*) Huggins, D. J., (2016). Studying the role of cooperative hydration in stabilizing folded protein states. *J. Struct. Biol. 196*, 394–406.

(*45*) Yuan, S., Filipek, S., Palczewski, K., and Vogel, H., (2014). Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. *Nat. Commun. 5*, 4733.

(*46*) Garczarek, F., and Gerwert, K., (2006). Functional waters in intraprotein proton transfer monitored by FTIR difference spectroscopy. *Nature 439*, 109–112.

(*47*) Sun, X., Ågren, H., and Tu, Y., (2014). Functional water molecules in rhodopsin activation. *J. Phys. Chem. B 118*, 10863–10873.

(*48*) Hofmann, K. P., Scheerer, P., Hildebrand, P. W., Choe, H. W., Park, J. H., Heck, M., and Ernst, O. P., (2009). A G protein-coupled receptor at work: the rhodopsin model. *Trends Biochem. Sci. 34*, 540–552.

(*49*) Pardo, L., Deupi, X., Dölker, N., López-Rodríguez, M. L., and Campillo, M., (2007). The role of internal water molecules in the structure and function of the rhodopsin family of G protein-coupled receptors. *ChemBioChem 8*, 19–24.

(*50*) Katritch, V., Fenalti, G., Abola, E. E., Roth, B. L., Cherezov, V., and Stevens, R. C., (2014). Allosteric sodium in class A GPCR signaling. *Trends Biochem. Sci. 39*, 233–244.

(*51*) Rahaman, O., Kalimeri, M., Melchionna, S., Henin, J., and Sterpone, F., (2015). Role of Internal Water on Protein Thermal Stability: The Case of Homologous G Domains. *J. Phys. Chem. B 119*, 8939–8949.

(*52*) Chakraborty, D., Taly, A., and Sterpone, F., (2015). Stay Wet, Stay Stable? How Internal Water Helps the Stability of Thermophilic Proteins. *J. Phys. Chem. B 119*, 12760–12770.

(*53*) Levitt, M., and Park, B. H., (1993). Water: now you see it, now you don't. *Structure 1*, 223–226.

(*54*) Karplus, P. A., and Faerman, C., (1994). Ordered water in macromolecular structure. *Curr. Opin. Struct. Biol. 4*, 770–776.

(*55*) Raschke, T. M., (2006). Water structure and interactions with protein surfaces. *Curr. Opin. Struct. Biol. 16*, 152–159.

(*56*) de Beer, S., Vermeulen, N., and Oostenbrink, C., (2010). The Role of Water Molecules in Computational Drug Design. *Curr. Top. Med. Chem. 10*, 55–66.

(*57*) Poornima, C. S., and Dean, P. M., (1995). Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. *J. Comput. Aided. Mol. Des. 9*, 521–531.

(*58*) Lu, Y., Wang, R., Yang, C. Y., and Wang, S., (2007). Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J. Chem. Inf. Model. 47*, 668–675.

(*59*) Williams, M. A., Goodfellow, J. M., and Thornton, J. M., (1994). Buried waters and internal cavities in monomeric proteins. *Protein Sci. 3*, 1224–1235.

(*60*) Ernst, J. A., Clubb, R. T., Zhou, H. X., Gronenborn, A. M., and Clore, G. M., (1995). Demonstration of positionally disordered water within a protein hydrophobic cavity by NMR. *Science 267*, 1813–1817.

(*61*) Vaitheeswaran, S., Yin, H., Rasaiah, J. C., and Hummer, G., (2004). Water clusters in nonpolar cavities. *Proc. Natl. Acad. Sci. U. S. A. 101*, 17002–17005.

(*62*) Haider, K., Wickstrom, L., Ramsey, S., Gilson, M. K., and Kurtzman, T., (2016). Enthalpic Breakdown of Water Structure on Protein Active-Site Surfaces. *J. Phys. Chem. B 120*, 8743–8756.

(*63*) Ahmed, M. H., Spyrakis, F., Cozzini, P., Tripathi, P. K., Mozzarelli, A., Scarsdale, J. N., Safo, M. A., and Kellogg, G. E., (2011). Bound water at protein-protein interfaces: Partners, roles and hydrophobic bubbles as a conserved motif. *PLoS One 6*, e24712.

(*64*) McMahon, C., Studer, S. M., Clendinen, C., Dann, G. P., Jeffrey, P. D., and Hughson, F. M., (2012). The structure of Sec12 implicates potassium ion coordination in Sar1 activation. *J. Biol. Chem. 287*, 43599–43606.

(*65*) Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E., (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem. 25*, 1605–1612.

(*66*) Park, S., and Saven, J. G., (2005). Statistical and molecular dynamics studies of buried waters in globular proteins. *Proteins Struct. Funct. Genet. 60*, 450–463.

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

(*68*) Biela, A., Khayat, M., Tan, H., Kong, J., Heine, A., Hangauer, D., and Klebe, G., (2012). Impact of ligand and protein desolvation on ligand binding to the S1 pocket of thrombin. *J. Mol. Biol. 418*, 350–366.

(*69*) Dunitz, J. D., (1994). The Entropic Cost of Bound Water in Crystals and Biomolecules. *Science (80-.). 264*, 670–670.

(*70*) Li, Z., and Lazaridis, T., (2007). Water at biomolecular binding interfaces. *Phys. Chem. Chem. Phys. 9*, 573–581.

(*71*) Cooper, A., (2005). Heat capacity effects in protein folding and ligand binding: A re-evaluation of the role of water in biomolecular thermodynamics. *Biophys. Chem. 115*, 89–97.

(*72*) Holdgate, G., In *Methods Mol. Biol.* Roque, A. C. A., Ed.; Humana Press: Totowa, NJ, 2009, pp 101–33.

(*73*) Freire, E., In *Thermodyn. Kinet. Drug Bind.* Wiley-VCH Verlag GmbH & Co. KGaA: 2015, pp 1–13.

(*74*) Palencia, A., Camara-Artigas, A., Pisabarro, M. T., Martinez, J. C., and Luque, I., (2010). Role of interfacial water molecules in proline-rich ligand recognition by the Src homology 3 domain of Abl. *J. Biol. Chem. 285*, 2823–2833.

(*75*) Ghai, R., Falconer, R. J., and Collins, B. M., (2012). Applications of isothermal titration calorimetry in pure and applied research-survey of the literature from 2010. *J. Mol. Recognit. 25*, 32–52.

(*76*) Krimmer, S. G., and Klebe, G., (2015). Thermodynamics of protein-ligand interactions as a reference for computational analysis: How to assess accuracy, reliability and relevance of experimental data. *J. Comput. Aided. Mol. Des. 29*, 867–883.

(*77*) Freire, E., (2008). Do enthalpy and entropy distinguish first in class from best in class? *Drug Discov. Today 13*, 869–874.

(*78*) Lafont, V., Armstrong, A. A., Ohtaka, H., Kiso, Y., Mario Amzel, L., and Freire, E., (2007). Compensating enthalpic and entropic changes hinder binding affinity optimization. *Chem. Biol. Drug Des. 69*, 413–422.

(*79*) Breiten, B., Lockett, M. R., Sherman, W., Fujita, S., Al-Sayah, M., Lange, H., Bowers, C. M., Heroux, A., Krilov, G., and Whitesides, G. M., (2013). Water networks contribute to enthalpy/entropy compensation in protein-ligand binding. *J. Am. Chem. Soc. 135*, 15579–15584.

(*80*) Sharp, K., (2001). Entropy-enthalpy compensation: fact or artifact? *Protein Sci. 10*, 661–7.

(*81*) Chodera, J. D., and Mobley, D. L., (2013). Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. *Annu. Rev. Biophys. 42*, 121–142.

(*82*) Snyder, P. W., Mecinovic, J., Moustakas, D. T., Thomas, S. W., Harder, M., Mack, E. T., Lockett, M. R., Heroux, A., Sherman, W., and Whitesides, G. M., (2011). Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. *Proc. Natl. Acad. Sci. 108*, 17889–17894.

(*83*) Lockett, M. R., Lange, H., Breiten, B., Heroux, A., Sherman, W., Rappoport, D., Yau, P. O., Snyder, P. W., and Whitesides, G. M., (2013). The binding of benzoarylsulfonamide ligands to human carbonic anhydrase is insensitive to formal fluorination of the ligand. *Angew. Chemie - Int. Ed. 52*, 7714–7717.

(*84*) Fährrolfes, R., Bietz, S., Flachsenberg, F., Meyder, A., Nittinger, E., Otto, T., Volkamer, A., and Rarey, M., (2017). Proteins Plus: A web portal for structure analysis of macromolecules. *Nucleic Acids Res. 45*, W337–W343.

(*85*) Tanford, C., (1978). The hydrophobic effect and the organization of living matter. *Science (80-.). 200*, 1012–1018.

(*86*) Englert, L., Biela, A., Zayed, M., Heine, A., Hangauer, D., and Klebe, G., (2010). 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. 1800*, 1192–1202.

(*87*) Biela, A., Sielaff, F., Terwesten, F., Heine, A., Steinmetzer, T., and Klebe, G., (2012). Ligand binding stepwise disrupts water network in thrombin: Enthalpic and entropic changes reveal classical hydrophobic effect. *J. Med. Chem. 55*, 6094–6110.

(*88*) Biela, A., Nasief, N. N., Betz, M., Heine, A., Hangauer, D., and Klebe, G., (2013). Dissecting the hydrophobic effect on the molecular level: The role of water, enthalpy, and entropy in ligand binding to thermolysin. *Angew. Chemie - Int. Ed. 52*, 1822–1828.

(*89*) Eisenberg, D., and Mclachlan, A. D., (1986). Solvation energy in protein folding and binding. *Nature 319*, 199–203.

(*90*) Hermann, R. B., (1972). Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem. 76*, 2754–2759.

(*91*) Pace, C. N., (1992). Contribution of the hydrophobic effect to globular protein stability. *J. Mol. Biol. 226*, 29–35.

(*92*) Setny, P., Baron, R., and McCammon, J. A., (2010). How can hydrophobic association be enthalpy driven? *J. Chem. Theory Comput. 6*, 2866–2871.

(*93*) Setny, P., Baron, R., and McCammon, J. A., Comment on 'Molecular driving forces of the pocket-ligand hydrophobic association' by G. Graziano, Chem. Phys. Lett. 533 (2012) 95., 2013.

(*94*) Graziano, G., (2012). Molecular driving forces of the pocket-ligand hydrophobic association. *Chem. Phys. Lett. 533*, 95–99.

(*95*) Copeland, R. A., Pompliano, D. L., and Meek, T. D., (2006). Drug–target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov. 5*, 730–739.

(*96*) Tummino, P. J., and Copeland, R. A., (2008). Residence time of receptor - Ligand complexes and its effect on biological function. *Biochemistry 47*, 5481–5492.

(*97*) Pan, A. C., Borhani, D. W., Dror, R. O., and Shaw, D. E., (2013). Molecular determinants of drug-receptor binding kinetics. *Drug Discov. Today 18*, 667–673.

(*98*) Schuetz, D. A., et al. (2017). Kinetics for Drug Discovery: an industry-driven effort to target drug residence time. *Drug Discov. Today 22*, 896–911.

(*99*) Schmidtke, P., Javier Luque, F., Murray, J. B., and Barril, X., (2011). Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. *J. Am. Chem. Soc. 133*, 18903–18910.

(*100*) Bortolato, A., Tehan, B. G., Bodnarchuk, M. S., Essex, J. W., and Mason, J. S., (2013). Water network perturbation in ligand binding: Adenosine A2A antagonists as a case study. *J. Chem. Inf. Model. 53*, 1700–1713.

(*101*) Buch, I., Giorgino, T., and De Fabritiis, G., (2011). Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. 108*, 10184–10189.

(*102*) Decherchi, S., Berteotti, A., Bottegoni, G., Rocchia, W., and Cavalli, A., (2015). The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nat. Commun. 6*, 6155.

(*103*) Deganutti, G., and Moro, S., (2017). Estimation of kinetic and thermodynamic ligand-binding parameters using computational strategies. *Future Med. Chem. 9*, 507–523.

(*104*) Zwanzig, R. W., (1954). High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys. 22*, 1420–1426.

(*105*) Price, M. L., and Jorgensen, W. L., (2000). Analysis of binding affinities for celecoxib analogues with COX-1 and COX-2 from combined docking and Monte Carlo simulations and insight into the COX-2/COX-1 selectivity. *J. Am. Chem. Soc. 122*, 9455–9466.

(*106*) Price, D. J., and Jorgensen, W. L., (2000). Computational binding studies of human pp60(c-src) SH2 domain with a series of nonpeptide, phosphophenyl-containing ligands. *Bioorganic Med. Chem. Lett. 10*, 2067–2070.

(*107*) Price, D. J., and Jorgensen, W. L., (2001). Improved convergence of binding affinities with free energy perturbation: Application to nonpeptide ligands with pp60src SH2 domain. *J. Comput. Aided. Mol. Des. 15*, 681–695.

(*108*) Barril, X., Orozco, M., and Luque, F. J., (1999). Predicting relative binding free energies of tacrine-huperzine A hybrids as inhibitors of acetylcholinesterase. *J. Med. Chem. 42*, 5110–5119.

(*109*) Oostenbrink, B. C., Pitera, J. W., Van Lipzig, M. M. H., Meerman, J. H. N., and Van Gunsteren, W. F., (2000). Simulations of the estrogen receptor ligand-binding domain: Affinity of natural ligands and xenoestrogens. *J. Med. Chem. 43*, 4594–4605.

(*110*) Pearlman, D. A., and Charifson, P. S., (2001). Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *J. Med. Chem. 44*, 3417–3423.

(*111*) Lazaridis, T., (1998). Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. *J. Phys. Chem. B 102*, 3531–3541.

(*112*) Lazaridis, T., (1998). Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids. *J. Phys. Chem. B 102*, 3542–3550.

(*113*) Huggins, D. J., and Payne, M. C., (2013). Assessing the accuracy of inhomogeneous fluid solvation theory in predicting hydration free energies of simple solutes. *J. Phys. Chem. B 117*, 8232–8244.

(*114*) Henchman, R. H., (2007). Free energy of liquid water from a computer simulation via cell theory. *J. Chem. Phys. 126*, 064504.

(*115*) Fogolari, F., Brigo, A., and Molinari, H., (2002). The Poisson-Boltzmann equation for biomolecular electrostatics: A tool for structural biology. *J. Mol. Recognit. 15*, 377–392.

(*116*) Baker, N. A., (2005). Improving implicit solvent simulations: A Poisson-centric view. *Curr. Opin. Struct. Biol. 15*, 137–143.

(*117*) Grochowski, P., and Trylska, J., (2008). Review: Continuum molecular electrostatics, salt effects, and counterion binding - A review of the Poisson-Boltzmann theory and its modifications. *Biopolymers 89*, 93–113.

(*118*) Bashford, D., and Case, D. A., (2000). Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem. 51*, 129–152.

(*119*) Chen, J., Brooks, C. L., and Khandogin, J., (2008). Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin. Struct. Biol. 18*, 140–148.

(*120*) Friesner, R. A., Abel, R., Goldfeld, D. A., Miller, E. B., and Murrett, C. S., (2013). Computational methods for high resolution prediction and refinement of protein structures. *Curr. Opin. Struct. Biol. 23*, 177–184.

(*121*) Kleinjung, J., and Fraternali, F., (2014). Design and application of implicit solvent models in biomolecular simulations. *Curr. Opin. Struct. Biol. 25*, 126–134.

(*122*) Gallicchio, E., Paris, K., and Levy, R. M., (2009). The AGBNP2 implicit solvation model. *J. Chem. Theory Comput. 5*, 2544–2564.

(*123*) Ren, P., Chun, J., Thomas, D. G., Schnieders, M. J., Zhang, J., and Baker, N. a., (2013). Biomolecular electrostatics and solvation: a computational perspective. *Q. Rev. Biophys. 45*, 427–491.

(*124*) Decherchi, S., Masetti, M., Vyalov, I., and Rocchia, W., (2015). Implicit solvent methods for free energy estimation. *Eur. J. Med. Chem. 91*, 27–42.

(*125*) Laage, D., Elsaesser, T., and Hynes, J. T., (2017). Water Dynamics in the Hydration Shells of Biomolecules. *Chem. Rev. 117*, 10694–10725.

(*126*) Wang, J., Cieplak, P., and Kollman, P. A., (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem. 21*, 1049–1074.

(*127*) Oostenbrink, C., Villa, A., Mark, A. E., and Van Gunsteren, W. F., (2004). 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. 25*, 1656–1676.

(*128*) Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J., (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc. 118*, 11225–11236.

(*129*) Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. V., and Hermans, J., In *Intermol. Forces. Jerusalem Symp. Quantum Chem. Biochem.* Pullman, B., Ed.; Springer Netherlands: Dordrecht, 1981, pp 331–342.

(*130*) Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P., (1987). The missing term in effective pair potentials. *J. Phys. Chem. 91*, 6269–6271.

(*131*) Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L., (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys. 79*, 926–935.

(*132*) Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., and Head-Gordon, T., (2004). Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys. 120*, 9665–9678.

(*133*) Hess, B., and van der Vegt, N. F. A., (2006). Hydration Thermodynamic Properties of Amino Acid Analogues: A Systematic Comparison of Biomolecular Force Fields and Water Models. *J. Phys. Chem. B 110*, 17616–17626.

(*134*) Onufriev, A. V., and Izadi, S., (2017). Water models for biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* e1347.

(*135*) Raymer, M. L., Sanschagrin, P. C., Punch, W. F., Venkataraman, S., Goodman, E. D., and Kuhn, L. a., (1997). Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol. 265*, 445–64.

(*136*) Sanschagrin, P. C., and Kuhn, L. a., (1998). Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity. *Protein Sci. 7*, 2054–64.

(*137*) García-Sosa, A. T., Mancera, R. L., and Dean, P. M., (2003). 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. 9*, 172–182.

(*138*) Patel, H., Grüning, B. A., Günther, S., and Merfort, I., (2014). PyWATER: a PyMOL plug-in to find conserved water molecules in proteins by clustering. *Bioinformatics 30*, 2978–2980.

(*139*) Kellogg, G. E., and Chen, D. L., (2004). The importance of being exhaustive. Optimization of bridging structural water molecules and water networks in models of biological systems. *Chem. Biodivers. 1*, 98–105.

(*140*) Amadasi, A., Surface, J. A., Spyrakis, F., Cozzini, P., Mozzarelli, A., and Kellogg, G. E., (2008). Robust classification of "relevant" water molecules in putative protein binding sites. *J. Med. Chem. 51*, 1063–1067.

(*141*) Kellogg, G. E., and Abraham, D. J., (2000). Hydrophobicity: Is LogP(o/w) more than the sum of its parts? *Eur. J. Med. Chem. 35*, 651–661.

(*142*) Pitt, W. R., and Goodfellow, J. M., (1991). Modelling of solvent positions around polar groups in proteins. *Protein Eng. Des. Sel. 4*, 531–537.

(*143*) Pitt, W. R., Murray-Rust, J., and Goodfellow, J. M., (1993). AQUARIUS2: Knowledge-based modeling of solvent sites around proteins. *J. Comput. Chem. 14*, 1007–1018.

(*144*) Bui, H. H., Schiewe, A. J., and Haworth, I. S., (2007). WATGEN: An algorithm for modeling water networks at protein-protein interfaces. *J. Comput. Chem. 28*, 2241–2251.

(*145*) Rossato, G., Ernst, B., Vedani, A., and Smieško, M., (2011). AcquaAlta: A directional approach to the solvation of ligand-protein complexes. *J. Chem. Inf. Model. 51*, 1867–1881.

(*146*) Ross, G. A., Morris, G. M., and Biggin, P. C., (2012). Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLoS One 7*, ed. by Csermely, P., e32036.

(*147*) Xiao, W., He, Z., Sun, M., Li, S., and Li, H., (2017). Statistical Analysis, Investigation, and Prediction of the Water Positions in the Binding Sites of Proteins. *J. Chem. Inf. Model. 57*, 1517–1528.

(*148*) Kellogg, G. E. G., Fornabaio, M., Chen, D. L., and Abraham, D. J., (2003). New application design for a 3D hydropathic map-based search for potential water molecules bridging between protein and ligand. *Internet Electron J. Mol. Des. 4*, 0–14.

(*149*) Schymkowitz, J. W. H., Rousseau, F., Martins, I. C., Ferkinghoff-Borg, J., Stricher, F., and Serrano, L., (2005). Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl. Acad. Sci. U. S. A. 102*, 10147–10152.

(*150*) Morozenko, A., and Stuchebrukhov, A. A., (2016). Dowser++, a new method of hydrating protein structures. *Proteins Struct. Funct. Bioinforma. 84*, 1347–1357.

(*151*) Allen, F. H., (2002). The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr. Sect. B Struct. Sci. 58*, 380–388.

(*152*) Goodsell, D. S., and Olson, A. J., (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins Struct. Funct. Bioinforma. 8*, 195–202.

(*153*) Trott, O., and Olson, A. J., (2010). Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem. 31*, 455–461.

(*154*) Morozenko, A., Leontyev, I. V., and Stuchebrukhov, A. A., (2014). Dipole moment and binding energy of water in proteins from crystallographic analysis. *J. Chem. Theory Comput. 10*, 4618–4623.

(*155*) Goodford, P. J., (1985). A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem. 28*, 849–857.

(*156*) Wade, R., Clark, K., and Goodford, P., (1993). Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem. 36*, 140–147.

(*157*) Wade, R. C., and Goodford, P. J., (1993). Further Development of Hydrogen Bond Functions for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. 2. Ligand Probe Groups with the Ability To Form More Than Two Hydrogen Bonds. *J. Med. Chem. 36*, 148–156.

(*158*) Miranker, A., and Karplus, M., (1991). Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins Struct. Funct. Bioinforma. 11*, 29–34.

(*159*) Bitetti-Putzer, R., Joseph-McCarthy, D., Hogle, J. M., and Karplus, M., (2001). Functional group placement in protein binding sites: A comparison of GRID and MCSS. *J. Comput. Aided. Mol. Des. 15*, 935–960.

(*160*) Evensen, E., Joseph-McCarthy, D., Weiss, G. A., Schreiber, S. L., and Karplus, M., (2007). Ligand design by a combinatorial approach based on modeling and experiment: Application to HLA-DR4. *J. Comput. Aided. Mol. Des. 21*, 395–418.

(*161*) Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F., and Mason, J. S., (2007). A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): Theory and application. *J. Chem. Inf. Model. 47*, 279–294.

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

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

(*164*) Zheng, M., Li, Y., Xiong, B., Jiang, H., and Shen, J., (2013). Water PMF for predicting the properties of water molecules in protein binding site. *J. Comput. Chem. 34*, 583–592.

(*165*) Bayden, A. S., Moustakas, D. T., Joseph-McCarthy, D., and Lamb, M. L., (2015). Evaluating Free Energies of Binding and Conservation of Crystallographic Waters Using SZMAP. *J. Chem. Inf. Model. 55*, 1552–1565.

(*166*) Setny, P., and Zacharias, M., (2010). Hydration in discrete water. A mean field, cellular automata based approach to calculating hydration free energies. *J. Phys. Chem. B 114*, 8667–8675.

(*167*) Setny, P., (2015). Hydration in discrete water (II): From neutral to charged solutes. *J. Phys. Chem. B 119*, 5970–5978.

(*168*) Setny, P., (2015). Prediction of Water Binding to Protein Hydration Sites with a Discrete, Semiexplicit Solvent Model. *J. Chem. Theory Comput. 11*, 5961–5972.

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

(*170*) Abel, R., Young, T., Farid, R., Berne, B. J., and Friesner, R. A., (2008). Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc. 130*, 2817–2831.

(*171*) Li, Z., and Lazaridis, T., (2012). Computing the thermodynamic contributions of interfacial water. *Methods Mol. Biol. 819*, 393–404.

(*172*) Cui, G., Swails, J. M., and Manas, E. S., (2013). SPAM: A simple approach for profiling bound water molecules. *J. Chem. Theory Comput. 9*, 5539–5549.

(*173*) Hu, B., and Lill, M. A., (2014). WATsite: Hydration site prediction program with PyMOL interface. *J. Comput. Chem. 35*, 1255–1260.

(*174*) Yang, Y., Hu, B., and Lill, M. A., (2017). WATsite2.0 with PyMOL plugin: Hydration site prediction and visualization. *Methods Mol. Biol. 1611*, 123–134.

(*175*) Gerogiokas, G., Calabro, G., Henchman, R. H., Southey, M. W., Law, R. J., and Michel, J., (2014). Prediction of small molecule hydration thermodynamics with grid cell theory. *J. Chem. Theory Comput. 10*, 35–48.

(*176*) Michel, J., Henchman, R. H., Gerogiokas, G., Southey, M. W. Y., Mazanetz, M. P., and Law, R. J., (2014). Evaluation of host-guest binding thermodynamics of model cavities with grid cell theory. *J. Chem. Theory Comput. 10*, 4055–4068.

(*177*) Gerogiokas, G., Southey, M. W. Y., Mazanetz, M. P., Hefeitz, A., Bodkin, M., Law, R. J., and Michel, J., (2015). Evaluation of water displacement energetics in protein binding sites with grid cell theory. *Phys. Chem. Chem. Phys. 17*, 8416–8426.

(*178*) López, E. D., Arcon, J. P., Gauto, D. F., Petruk, A. A., Modenutti, C. P., Dumas, V. G., Marti, M. A., and Turjanski, A. G., (2015). WATCLUST: A tool for improving the design of drugs based on protein-water interactions. *Bioinformatics 31*, 3697–3699.

(*179*) Zia, S. R., Gaspari, R., Decherchi, S., and Rocchia, W., (2016). Probing Hydration Patterns in Class-A GPCRs via Biased MD: The A2A Receptor. *J. Chem. Theory Comput. 12*, 6049–6061.

(*180*) Nguyen, C. N., Young, T. K., and Gilson, M. K., (2012). Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *J. Chem. Phys. 137*, 044101.

(*181*) Nguyen, C. N., Cruz, A., Gilson, M. K., and Kurtzman, T., (2014). Thermodynamics of Water in an Enzyme Active Site : Grid-Based Hydration Analysis of Coagulation Factor Xa. *J. Chem. Theory Comput. 10*, 2769–2780.

(*182*) Ghanakota, P., and Carlson, H. A., (2016). Moving beyond Active-Site Detection: MixMD Applied to Allosteric Systems. *J. Phys. Chem. B 120*, 8685–8695.

(*183*) Graham, S. E., Smith, R. D., and Carlson, H. A., (2018). Predicting Displaceable Water Sites Using Mixed-Solvent Molecular Dynamics. *J. Chem. Inf. Model. 58*, acs.jcim.7b00268.

(*184*) Woods, C. J., Essex, J. W., and King, M. A., (2003). Enhanced Configurational Sampling in Binding Free-Energy Calculations. *J Phys Chem B 107*, 13711–13718.

(*185*) Woods, C. J., Essex, J. W., and King, M. A., (2003). The Development of Replica-Exchange-Based Free-Energy Methods. *J. Phys. Chem. B 107*, 13703–13710.

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

(*187*) Hamelberg, D., and McCammon, J. A., (2004). Standard free energy of releasing a localized water molecule from the binding pockets of proteins: Double-decoupling method. *J. Am. Chem. Soc. 126*, 7683–7689.

(*188*) Barillari, C., Taylor, J., Viner, R., and Essex, J. W., (2007). Classification of water molecules in protein binding sites. *J. Am. Chem. Soc. 129*, 2577–2587.

(*189*) Michel, J., Tirado-Rives, J., and Jorgensen, W. L., (2009). Prediction of the water content in protein binding sites. *J. Phys. Chem. B 113*, 13337–13346.

(*190*) Rakhmanov, S. V., and Makeev, V. J., (2007). Atomic hydration potentials using a Monte Carlo Reference State (MCRS) for protein solvation modeling. *BMC Struct. Biol. 7*, 19.

(*191*) Adams, D. J., (1974). Chemical potential of hard-sphere fluids by monte carlo methods. *Mol. Phys. 28*, 1241–1252.

(*192*) Adams, D., (1975). Grand canonical ensemble Monte Carlo for a Lennard-Jones fluid. *Mol. Phys. 29*, 307–311.

(*193*) Woo, H.-J., Dinner, A. R., and Roux, B., (2004). Grand canonical Monte Carlo simulations of water in protein environments. *J. Chem. Phys. 121*, 6392–6400.

(*194*) Ross, G. A., Bodnarchuk, M. S., and Essex, J. W., (2015). Water Sites, Networks, and Free Energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc. 137*, 14930–14943.

(*195*) Afanasyeva, A., Izmailov, S., Grigoriev, M., and Petukhov, M., (2015). AquaBridge: A novel method for systematic search of structural water molecules within the protein active sites. *J. Comput. Chem. 36*, 1973–1977.

(*196*) Sleigh, S. H., Seavers, P. R., Wilkinson, A. J., Ladbury, J. E., and Tame, J. R., (1999). Crystallographic and calorimetric analysis of peptide binding to OppA protein. *J. Mol. Biol. 291*, 393–415.

(*197*) Lensink, M. F., et al. (2014). Blind prediction of interfacial water positions in CAPRI. *Proteins Struct. Funct. Bioinforma. 82*, 620–632.

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

(*199*) Graves, A. P., Wall, I. D., Edge, C. M., Woolven, J. M., Cui, G., Le Gall, A., Hong, X., Raha, K., and Manas, E. S., (2017). A Perspective on Water Site Prediction Methods for Structure Based Drug Design. *Curr. Top. Med. Chem. 17*, 2599–2616.

(*200*) Bodnarchuk, M. S., Viner, R., Michel, J., and Essex, J. W., (2014). Strategies to calculate water binding free energies in protein-ligand complexes. *J. Chem. Inf. Model. 54*, 1623–1633.

(*201*) Mason, J. S., Bortolato, A., Congreve, M., and Marshall, F. H., (2012). New insights from structural biology into the druggability of G protein-coupled receptors. *Trends Pharmacol. Sci. 33*, 249–260.

(*202*) Mason, J. S., Bortolato, A., Weiss, D. R., Deflorian, F., Tehan, B., and Marshall, F. H., (2013). High end GPCR design: crafted ligand design and druggability analysis using protein structure, lipophilic hotspots and explicit water networks. *Silico Pharmacol. 1*, 23.

(*203*) Bucher, D., Stouten, P., and Triballeau, N., (2018). Shedding Light on Important Waters for Drug Design: Simulations versus Grid-Based Methods. *J. Chem. Inf. Model.* acs.jcim.7b00642.

(*204*) Verdonk, M. L., Chessari, G., Cole, J. C., Hartshorn, M. J., Murray, C. W., Nissink, J. W. M., Taylor, R. D., and Taylor, R., (2005). Modeling water molecules in protein-ligand docking using GOLD. *J. Med. Chem. 48*, 6504–6515.

(*205*) 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., and Shenkin, P. S., (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem. 47*, 1739–1749.

(*206*) Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C., and Mainz, D. T., (2006). Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem. 49*, 6177–6196.

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

(*208*) Rarey, M., Kramer, B., Lengauer, T., and Klebe, G., (1996). A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol. 261*, 470–489.

(*209*) Rarey, M., Kramer, B., and Lengauer, T., (1999). The particle concept: Placing discrete water molecules during protein- ligand docking predictions. *Proteins Struct. Funct. Genet. 34*, 17–28.

(*210*) Corbeil, C. R., Englebienne, P., and Moitessier, N., (2007). Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model. 47*, 435–449.

(*211*) Corbeil, C. R., and Moitessier, N., (2009). Docking ligands into flexible and solvated macromolecules. 3. Impact of input ligand conformation, protein flexibility, and water molecules on the accuracy of docking programs. *J. Chem. Inf. Model. 49*, 997–1009.

(*212*) Therrien, E., Weill, N., Tomberg, A., Corbeil, C. R., Lee, D., and Moitessier, N., (2014). Docking ligands into flexible and solvated macromolecules. 7. Impact of protein flexibility and water molecules on docking-based virtual screening accuracy. *J. Chem. Inf. Model. 54*, 3198–3210.

(*213*) Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J., (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem. 19*, 1639–1662.

(*214*) Morris, G. M., Ruth, H., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J., (2009). Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem. 30*, 2785–2791.

(*215*) Uehara, S., and Tanaka, S., (2016). AutoDock-GIST: Incorporating thermodynamics of active-site water into scoring function for accurate protein-ligand docking. *Molecules 21*, 1604.

(*216*) Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K., (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem. 55*, 6582–6594.

(*217*) Huey, R., Morris, G. M., Olson, A. J., and Goodsell, D. S., (2007). Software news and update a semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem. 28*, 1145–1152.

(*218*) Forli, S., and Olson, A. J., (2012). A force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *J. Med. Chem. 55*, 623–638.

(*219*) Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., and Olson, A. J., (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc. 11*, 905–919.

(*220*) Moitessier, N., Westhof, E., and Hanessian, S., (2006). Docking of aminoglycosides to hydrated and flexible RNA. *J. Med. Chem. 49*, 1023–1033.

(*221*) Balius, T. E., Fischer, M., Stein, R. M., Adler, T. B., Nguyen, C. N., Cruz, A., Gilson, M. K., Kurtzman, T., and Shoichet, B. K., (2017). Testing inhomogeneous solvation theory in structure-based ligand discovery. *Proc. Natl. Acad. Sci. 114*, 201703287.

(*222*) Huang, N., and Shoichet, B. K., (2008). Exploiting ordered waters in molecular docking. *J. Med. Chem. 51*, 4862–4865.

(*223*) Murphy, R. B., Repasky, M. P., Greenwood, J. R., Tubert-Brohman, I., Jerome, S., Annabhimoju, R., Boyles, N. A., Schmitz, C. D., Abel, R., Farid, R., and Friesner, R. A., (2016). WScore: A Flexible and Accurate Treatment of Explicit Water Molecules in Ligand-Receptor Docking. *J. Med. Chem. 59*, 4364–4384.

(*224*) Roberts, B. C., and Mancera, R. L., (2008). Ligand - protein docking with water molecules. *J. Chem. Inf. Model. 48*, 397–408.

(*225*) Thilagavathi, R., and Mancera, R. L., (2010). Ligand-protein cross-docking with water molecules. *J. Chem. Inf. Model. 50*, 415–421.

(*226*) Kumar, A., and Zhang, K. Y. J., (2013). Investigation on the effect of key water molecules on docking performance in CSARdock exercise. *J. Chem. Inf. Model. 53*, 1880–1892.

(*227*) Li, L., Xu, W., and Lü, Q., (2015). Improving protein-ligand docking with flexible interfacial water molecules using SWRosettaLigand. *J. Mol. Model. 21*, 294.

(*228*) Pastor, M., Cruciani, G., and Watson, K. A., (1997). A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure-activity relationship analysis. *J. Med. Chem. 40*, 4089–102.

(*229*) Wang, T., and Wade, R. C., (2001). Comparative binding energy (COMBINE) analysis of influenza neuraminidase-inhibitor complexes. *J. Med. Chem. 44*, 961–971.

(*230*) Hussain, A., Melville, J. L., and Hirst, J. D., (2010). Molecular docking and QSAR of aplyronine A and analogues: Potent inhibitors of actin. *J. Comput. Aided. Mol. Des. 24*, 1–15.

(*231*) Taha, M. O., Habash, M., Al-Hadidi, Z., Al-Bakri, A., Younis, K., and Sisan, S., (2011). Docking-based comparative intermolecular contacts analysis as new 3-D QSAR concept for validating docking studies and in silico screening: NMT and GP inhibitors as case studies. *J. Chem. Inf. Model. 51*, 647–669.

(*232*) Brenk, R., Naerum, L., Grädler, U., Gerber, H. D., Garcia, G. A., Reuter, K., Stubbs, M. T., and Klebe, G., (2003). Virtual screening for submicromolar leads of tRNA-guanine transglycosylase based on a new unexpected binding mode detected by crystal structure analysis. *J. Med. Chem. 46*, 1133–1143.

(*233*) Lloyd, D. G., García-Sosa, A. T., Alberts, I. L., Todorov, N. P., and Mancera, R. L., (2004). The effect of tightly bound water molecules on the structural interpretation of ligand-derived pharmacophore models. *J. Comput. Aided. Mol. Des. 18*, 89–100.

(*234*) Blum, A. P., Lester, H. A., and Dougherty, D. A., (2010). Nicotinic pharmacophore: The pyridine N of nicotine and carbonyl of acetylcholine hydrogen bond across a subunit interface to a backbone NH. *Proc. Natl. Acad. Sci. 107*, 13206–13211.

(*235*) Giménez-Oya, V., Villacañas, Ã., Obiol-Pardo, C., Antolin-Llovera, M., Rubio-Martinez, J., and Imperial, S., (2011). Design of novel ligands of CDP-methylerythritol kinase by mimicking direct protein-protein and solvent-mediated interactions. *J. Mol. Recognit. 24*, 71–80.

(*236*) Liu, C., et al. (2005). 5-Cyanopyrimidine Derivatives as a Novel Class of Potent, Selective, and Orally Active Inhibitors of p38α MAP Kinase. *J. Med. Chem. 48*, 6261–6270.

(*237*) Baum, B., Muley, L., Heine, A., Smolinski, M., Hangauer, D., and Klebe, G., (2009). Think Twice: Understanding the High Potency of Bis(phenyl)methane Inhibitors of Thrombin. *J. Mol. Biol. 391*, 552–564.

(*238*) Huang, P. P., Randolph, J. T., Klein, L. L., Vasavanonda, S., Dekhtyar, T., Stoll, V. S., and Kempf, D. J., (2004). Synthesis and antiviral activity of P1 arylsulfonamide azacyclic urea HIV protease inhibitors. *Bioorganic Med. Chem. Lett. 14*, 4075–4078.

(*239*) Andrews, S. P., Mason, J. S., Hurrell, E., and Congreve, M., (2014). Structure-based drug design of chromone antagonists of the adenosine A2A receptor. *Medchemcomm 5*, 571–575.

(*240*) Spyrakis, F., Ahmed, M. H., Bayden, A. S., Cozzini, P., Mozzarelli, A., and Kellogg, G. E., (2017). The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem. 60*, 6781–6828.

(*241*) Lundqvist, T., Rice, J., Hodge, C. N., Basarab, G. S., Pierce, J., and Lindqvist, Y., (1994). Crystal structure of scytalone dehydratase - a disease determinant of the rice pathogen, Magnaporthe grisea. *Structure 2*, 937–944.

(*242*) Biela, A., Betz, M., Heine, A., and Klebe, G., (2012). Water Makes the Difference: Rearrangement of Water Solvation Layer Triggers Non-additivity of Functional Group Contributions in Protein-Ligand Binding. *ChemMedChem 7*, 1423–1434.

(*243*) Rondelet, G., Dal Maso, T., Willems, L., and Wouters, J., (2016). Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B. *J. Struct. Biol. 194*, 357–367.

(*244*) Ge, Y. Z., Pu, M. T., Gowher, H., Wu, H. P., Ding, J. P., Jeltsch, A., and Xu, G. L., (2004). Chromatin targeting of de novo DNA methyltransferases by the PWWP domain. *J. Biol. Chem. 279*, 25447–25454.

(*245*) Salie, Z. L., Kirby, K. A., Michailidis, E., Marchand, B., Singh, K., Rohan, L. C., Kodama, E. N., Mitsuya, H., Parniak, M. A., and Sarafianos, S. G., (2016). Structural basis of HIV inhibition by translocation-defective RT inhibitor 4'-ethynyl-2-fluoro-2'-deoxyadenosine (EFdA). *Proc. Natl. Acad. Sci. U. S. A. 113*, 9274–9.

(*246*) Pujadas, G., and Palau, J., (2001). Molecular mimicry of substrate oxygen atoms by water molecules in the beta-amylase active site. *Protein Sci. 10*, 1645–1657.

(*247*) Levinson, N. M., and Boxer, S. G., (2013). A conserved water-mediated hydrogen bond network defines bosutinib's kinase selectivity. *Nat. Chem. Biol. 10*, 127–132.

(*248*) Baldwin, E. T., Bhat, T. N., Gulnik, S., Liu, B., Topol, I. A., Kiso, Y., Mimoto, T., Mitsuya, H., and Erickson, J. W., (1995). Structure of HIV-1 protease with KNI-272, a tight-binding transition-state analog containing allophenylnorstatine. *Structure 3*, 581–590.

(*249*) Cappel, D., Sherman, W., and Beuming, T., (2017). Calculating Water Thermodynamics in the Binding Site of Proteins – Applications of WaterMap to Drug Discovery. *Curr. Top. Med. Chem. 17*, 1–1.

(*250*) Bingham, R. J., Findlay, J. B. C., Hsieh, S.-Y., Kalverda, A. P., Kjellberg, A., Perazzolo, C., Phillips, S. E. V., Seshadri, K., Trinh, C. H., Turnbull, W. B., Bodenhausen, G., and Homans, S. W., (2004). Thermodynamics of binding of 2-methoxy-3-isopropylpyrazine and 2-methoxy-3-isobutylpyrazine to the major urinary protein. *J. Am. Chem. Soc. 126*, 1675–81.

(*251*) Haider, K., and Huggins, D. J., (2013). Combining solvent thermodynamic profiles with functionality maps of the Hsp90 binding site to predict the displacement of water molecules. *J. Chem. Inf. Model. 53*, 2571–2586.

(*252*) Li, Z., and Lazaridis, T., (2003). Thermodynamic contributions of the ordered water molecule in HIV-1 protease. *J. Am. Chem. Soc. 125*, 6636–6637.

(*253*) García-Sosa, A. T., and Mancera, R. L., (2010). Free energy calculations of mutations involving a tightly bound water molecule and ligand substitutions in a ligand-protein complex. *Mol. Inform. 29*, 589–600.

(*254*) Higgs, C., Beuming, T., and Sherman, W., (2010). Hydration site thermodynamics explain SARs for triazolylpurines analogues binding to the A2A receptor. *ACS Med. Chem. Lett. 1*, 160–164.

(*255*) Kadirvelraj, R., Foley, B. L., Dyekjær, J. D., and Woods, R. J., (2008). Involvement of water in carbohydrate-protein binding: Concanavalin A revisited. *J. Am. Chem. Soc. 130*, 16933–16942.

(*256*) Clarke, C., Woods, R. J., Gluska, J., Cooper, A., Nutley, M. A., and Boons, G.-j., (2001). Involvement of Water in Carbohydrate - Protein Binding. *J. Am. Chem. Soc.* 12238–12247.

(*257*) García-Sosa, A. T., (2013). 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. 53*, 1388–1405.

(*258*) Zheng, H., Hou, J., Zimmerman, M. D., Wlodawer, A., and Minor, W., (2014). The future of crystallography in drug discovery. *Expert Opin. Drug Discov. 9*, 125–137.

(*259*) Lysozyme crystals observed through polarizing filter in Nikon SMZ800 microscope. By Ivan Taavi, CC BY-SA 3.0. https://commons.wikimedia.org/w/index.php?curid=22143196.

(*260*) Diffraction image of protein crystal. Hen egg lysozyme P43212, a=b 7.872 nm, c 3.683 nm, α=β=γ 90°. By Del45, Public Domain. https://commons.wikimedia.org/w/index.php?curid=7923353.

(*261*) König, V., Pfeil, A., Braus, G. H., and Schneider, T. R., (2004). Substrate and Metal Complexes of 3-Deoxy-D-arabino-heptulosonate-7- phosphate Synthase from Saccharomyces cerevisiae Provide New Insights into the Catalytic Mechanism. *J. Mol. Biol. 337*, 675–690.

(*262*) Versées, W., Decanniere, K., Pellé, R., Depoorter, J., Brosens, E., Parkin, D. W., and Steyaert, J., (2001). Structure and function of a novel purine specific nucleoside hydrolase from Trypanosoma vivax. *J. Mol. Biol. 307*, 1363–1379.

(*263*) Bruno, I. J., Cole, J. C., Lommerse, J. P., Rowland, R. S., Taylor, R., and Verdonk, M. L., (1997). IsoStar: a library of information about nonbonded interactions. *J. Comput. Aided. Mol. Des. 11*, 525–537.

(*264*) Verdonk, M. L., Cole, J. C., and Taylor, R., (1999). SuperStar: A Knowledge-based Approach for Identifying Interaction Sites in Proteins. *J. Mol. Biol. 289*, 1093–1108.

(*265*) Verdonk, M. L., Cole, J. C., Watson, P., Gillet, V., and Willett, P., (2001). Superstar: improved knowledge-based interaction fields for protein binding sites11Edited by R. Huber. *J. Mol. Biol. 307*, 841–859.

(*266*) Urbaczek, S., Kolodzik, A., Fischer, J. R., Lippert, T., Heuser, S., Groth, I., Schulz-Gasch, T., and Rarey, M., (2011). NAOMI: On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model. 51*, 3199–3207.

(*267*) Urbaczek, S., Kolodzik, A., Groth, I., Heuser, S., and Rarey, M., (2013). Reading PDB: Perception of molecules from 3D atomic coordinates. *J. Chem. Inf. Model. 53*, 76–87.

(*268*) Urbaczek, S., Kolodzik, A., and Rarey, M., (2014). The valence state combination model: A generic framework for handling tautomers and protonation states. *J. Chem. Inf. Model. 54*, 756–766.

(*269*) Taylor, R., Kennard, O., and Versichel, W., (1984). The geometry of the N-H...O=C hydrogen bond. 3. Hydrogen-bond distances and angles. *Acta Crystallogr. Sect. B 40*, 280–288.

(*270*) Mills, J. E., and Dean, P. M., (1996). Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *J. Comput. Aided. Mol. Des. 10*, 607–622.

(*271*) Tóth, G., Bowers, S. G., Truong, A. P., and Probst, G., (2007). The role and significance of unconventional hydrogen bonds in small molecule recognition by biological receptors of pharmaceutical relevance. *Curr. Pharm. Des. 13*, 3476–3493.

(*272*) Horowitz, S., and Trievel, R. C., (2012). Carbon-oxygen hydrogen bonding in biological structure and function. *J. Biol. Chem. 287*, 41576–41582.

(*273*) Shing Ho, P., In *Halogen Bond. I. Top. Curr. Chem.* Metrangolo, P., and Resnati, G., Eds.; Topics in Current Chemistry, Vol. 358; Springer International Publishing: Cham, 2014, pp 241–276.

(*274*) Reiling, K. K., Endres, N. F., Dauber, D. S., Craik, C. S., and Stroud, R. M., (2002). Anisotropic dynamics of the JE-2147-HIV protease complex: Drug resistance and thermodynamic binding mode examined in a 1.09 A structure. *Biochemistry 41*, 4582–4594.

(*275*) Malamas, M. S., Erdei, J., Gunawan, I., Turner, J., Hu, Y., Wagner, E., Fan, K., Chopra, R., Olland, A., Bard, J., Jacobsen, S., Magolda, R. L., Pangalos, M., and Robichaud, A. J., (2010). Design and synthesis of 5,5-disubstituted aminohydantoins as potent and selective human β-secretase (BACE1) inhibitors. *J. Med. Chem. 53*, 1146–1158.

(*276*) Bietz, S., Urbaczek, S., Schulz, B., and Rarey, M., (2014). Protoss: A holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminform. 6*, 12.

(*277*) MacCuish, J. D., and MacCuish, N. E., *Clustering in Bioinformatics and Drug Discovery*; CRC Press: 2011, p 214.

(*278*) Nocedal, J., and Wright, S. J., In *Numer. Optim. Springer Ser. Oper. Res. Financ. Eng.* 2nd; Springer, New York: 2006; Chapter 6, pp 135–163.

(*279*) Loris, R., Tielker, D., Jaeger, K. E., and Wyns, L., (2003). Structural basis of carbohydrate recognition by the lectin LecB from Pseudomonas aeruginosa. *J. Mol. Biol. 331*, 861–870.

(*280*) Hoffmann, M. M., and Conradi, M. S., (1997). Are there hydrogen bonds in supercritical water? *J. Am. Chem. Soc. 119*, 3811–3817.

(*281*) Soper, A. K., Bruni, F., and Ricci, M. A., (1997). Site-site pair correlation functions of water from 25 to 400 °C: Revised analysis of new and old diffraction data. *J. Chem. Phys. 106*, 247–254.

(*282*) Wernet, P., (2004). The Structure of the First Coordination Shell in Liquid Water. *Science (80-.). 304*, 995–999.

(*283*) Eaves, J. D., Loparo, J. J., Fecko, C. J., Roberts, S. T., Tokmakoff, A., and Geissler, P. L., (2005). Hydrogen bonds in liquid water are broken only fleetingly. *Proc. Natl. Acad. Sci. 102*, 13019–13022.

(*284*) Barna, T. M., Khan, H., Bruce, N. C., Barsukov, I., Scrutton, N. S., and Moody, P. C., (2001). Crystal structure of pentaerythritol tetranitrate reductase: "Flipped" binding geometries for steroid substrates in different redox states of the enzyme. *J. Mol. Biol. 310*, 433–447.

(*285*) Parmeggiani, A., Krab, I. M., Okamura, S., Nielsen, R. C., Nyborg, J., and Nissen, P., (2006). Structural basis of the action of pulvomycin and GE2270 A on elongation factor Tu. *Biochemistry 45*, 6846–6857.

(*286*) Davis, T. L., Walker, J. R., Loppnau, P., Butler-Cole, C., Allali-Hassani, A., and Dhe-Paganon, S., (2008). Autoregulation by the Juxtamembrane Region of the Human Ephrin Receptor Tyrosine Kinase A3 (EphA3). *Structure 16*, 873–884.

(*287*) Schneider, N., Hindle, S., Lange, G., Klein, R., Albrecht, J., Briem, H., Beyer, K., Claußen, H., Gastreich, M., Lemmen, C., and Rarey, M., (2012). Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *J. Comput. Aided. Mol. Des. 26*, 701–723.

(*288*) Crawford, T. D., et al. (2016). Diving into the Water: Inducible Binding Conformations for BRD4, TAF1(2), BRD9, and CECR2 Bromodomains. *J. Med. Chem. 59*, 5391–5402.

(*289*) Pawson, T., and Nash, P., (2003). Assembly of cell regulatory systems through protein interaction domains. *Science 300*, 445–52.

(*290*) Nooren, I. M. A., and Thornton, J. M., (2003). Diversity of protein-protein interactions. *EMBO J. 22*, 3486–3492.

(*291*) Henrick, K., and Thornton, J. M., (1998). PQS: A protein quaternary structure file server. *Trends Biochem. Sci. 23*, 358–361.

(*292*) Ponstingl, H., Kabir, T., and Thornton, J. M., (2003). Automatic inference of protein quaternary structure from crystals. *J. Appl. Crystallogr. 36*, 1116–1122.

(*293*) Mintseris, J., and Weng, Z., (2003). Atomic Contact Vectors in Protein-Protein Recognition. *Proteins Struct. Funct. Genet. 53*, 629–639.

(*294*) Zhu, H., Domingues, F. S., Sommer, I., and Lengauer, T., (2006). NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics 7*, 27.

(*295*) Block, P., Paern, J., Hüllermeier, E., Sanschagrin, P., Sotriffer, C. A., and Klebe, G., (2006). Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins Struct. Funct. Genet. 65*, 607–622.

(*296*) Krissinel, E., and Henrick, K., (2007). Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol. 372*, 774–797.

(*297*) Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J., and Poupon, A., (2008). DiMoVo: A Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics 24*, 652–658.

(*298*) Schärer, M. A., Grütter, M. G., and Capitani, G., (2010). CRK: An evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. *Proteins Struct. Funct. Bioinforma. 78*, 2707–2713.

(*299*) Liu, Q., and Li, J., (2010). Propensity vectors of low-ASA residue pairs in the distinction of protein interactions. *Proteins Struct. Funct. Bioinforma. 78*, 589–602.

(*300*) Mitra, P., and Pal, D., (2011). Combining Bayes classification and point group symmetry under Boolean framework for enhanced protein quaternary structure inference. *Structure 19*, 304–312.

(*301*) Da Silva, F., Desaphy, J., Bret, G., and Rognan, D., (2015). IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein-Protein Interfaces. *J. Chem. Inf. Model. 55*, 2005–2014.

(*302*) Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F., Misc functions of the Department of Statistics (e1071), TU Wien., 2014.

(*303*) Arkin, M. R., Randal, M., DeLano, W. L., Hyde, J., Luong, T. N., Oslob, J. D., Raphael, D. R., Taylor, L., Wang, J., McDowell, R. S., Wells, J. A., and Braisted, A. C., (2003). Binding of small molecules to an adaptive protein-protein interface. *Proc. Natl. Acad. Sci. 100*, 1603–1608.

(*304*) Rickert, M., Wang, X., Boulanger, M. J., Goriatcheva, N., and Garcia, K. C., (2005). The Structure of Interleukin-2 Complexed with Its Alpha Receptor. *Science 308*, 1477–80.

(*305*) Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., and Robins, K., (2012). Engineering the third wave of biocatalysis. *Nature 485*, 185–194.

(*306*) Kaljunen, H., Chow, J., Streit, W. R., and Mueller-Dieckmann, J., (2014). Cloning, expression, purification and preliminary X-ray analysis of EstN2, a novel archaeal α/β-hydrolase from Candidatus Nitrososphaera gargensis. *Acta Crystallogr. Sect. FStructural Biol. Commun. 70*, 1394–1397.

(*307*) Schomburg, I., Chang, A., Placzek, S., Söhngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M., Grote, A., Scheer, M., and Schomburg, D., (2013). BRENDA in 2013: Integrated reactions, kinetic data, enzyme function data, improved disease classification: New options and contents in BRENDA. *Nucleic Acids Res. 41*, D764–72.

(*308*) Schulz, D., and Rentmeister, A., (2012). An enzyme-coupled high-throughput assay for screening RNA methyltransferase activity in E. Coli cell lysate. *RNA Biol. 9*, 577–586.

(*309*) Holstein, J. M., Anhäuser, L., and Rentmeister, A., (2016). Modifying the 5-Cap for Click Reactions of Eukaryotic mRNA and To Tune Translation Efficiency in Living Cells. *Angew. Chemie - Int. Ed. 55*, 10899–10903.

(*310*) Bietz, S., Methoden zur computergestützten Generierung und Aufbereitung von Strukturensembles für Proteinbindetaschen., Ph.D. Thesis, Universität Hamburg, 2016.

Bibliography of this Dissertation's Publications

D1 Schneider, N.; Volkamer, A.; **Nittinger, E.**; Rarey, M. Supporting Biocatalysis Research with Structural Bioinformatics. In Applied Biocatalysis: From Fundamental Science to Industrial Applications; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2016; pp 71–100.

D2 **Nittinger, E.**; Schneider, N.; Lange, G.; Rarey, M. Evidence of Water Molecules – A Statistical Evaluation of Water Molecules Based on Electron Density. J. Chem. Inf. Model. 2015, 55 (4): 771–783.

D3 Meyder, A.; **Nittinger, E.**; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. J. Chem. Inf. Model. 2017, 57 (10): 2437-2447.

D4 Inhester, T.; **Nittinger, E.**; Sommer, K.; Schmidt, P.; Bietz, S.; Rarey, M. *NAOMI*nova: Interactive Geometric Analysis of Noncovalent Interactions in Macromolecular Structures. J. Chem. Inf. Model. 2017, 57 (9): 2132-2142.

D5 **Nittinger, E.**; Inhester, T.; Bietz, S.; Meyder, A.; Schomburg, K.; Lange, G.; Klein, R.; Rarey, M. Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein-Ligand Interfaces. J. Med. Chem. 2017, 60 (10): 4245-4257.

D6 Fährrolfes, R.; Bietz, S.; Meyder, A.; **Nittinger, E.**; Otto, T.; Flachsenberg, F.; Volkamer, A.; Rarey, M. Proteins*Plus*: a web portal for structure analysis of macromolecules. Nucleic Acids Res. 2017, 45 (W1): W337-W343.

D7 Bietz, S.; Inhester, T.; Lauck, F.; Sommer, K.; von Behren, M. M.; Fährrolfes, R.; Flachsenberg, F.; Meyder, A.; **Nittinger, E.**; Otto, T.; Hilbig, M.; Schomburg, K. T.; Volkamer, A.; Rarey, M. From cheminformatics to structure-based design: Web services and desktop applications based on the NAOMI library. Journal of Biotechnology. J. Biotechnol. 2017, 261: 207-214.

D8 Schomburg, K. T.; **Nittinger, E.**; Meyder, A.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M. Prediction of protein mutation effects based on dehydration and hydrogen bonding – A large-scale study. Proteins Struct. Funct. Bioinforma. 2017, 85 (8): 1550-1566.

Additional Publications

E1 von Behren, M. M.; Bietz, S.; **Nittinger, E.**; Rarey, M. mRAISE: An Alternative Algorithmic Approach to Ligand-Based Virtual Screening. J. Comput. Aided. Mol. Des. 2016, 30 (8), 583–594.

Manuscripts in Preparation

P1 **Nittinger, E.**; Flachsenberg, F.; Bietz, S.; Lange, G.; Rarey, M. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. J. Chem. Inf. Model. 2018, *Accepted*.

P2 **Nittinger, E.**; Gibbons, P.; Eigenbrot, C.; Davies, D. R.; Maurer, B.; Yu, C. L.; Kiefer, J. R.; Kuglstatter, A.; Murray, J.; Ortwine, D. F.; Tang, Y.; Tsui, V. Water Molecules in Protein-Ligand Interfaces. Evaluation of Software Tools and SAR Comparison. J. Comput. Aided. Mol. Des., *Submitted for publication*.

Abbreviations

BRD	Bromodomain
BTK	Bruton's Tyrosine Kinase
CADD	computer-aided drug design
EDIA	electron density of individual atoms
EDIA$_m$	electron density of multiple atoms
FIP	free interaction point
FSI	free space identification
H/SC	enthalpy/entropy compensation
HYDE	hydrogen bonds and dehydration
IA	interaction
PDB	Protein Data Bank
PWP	potential water position
RSCC	real-space R correlation coefficient
RSR	real-space R factor
RSR-Z	normalized RSR
RSZD	real-space difference density Z score
RSZO	real-space observed density Z score
SAR	structure activity relationship
SBVS	structure-based virtual screening
WarPP	water placement procedure

A
Scientific Contributions

A.1 Publications in Scientific Journals

This section lists the author's publications in scientific journals and a book chapter. Additionally, the author's contributions are specified.

D1 Schneider, N.; Volkamer, A.; **Nittinger, E.**; Rarey, M. Supporting Biocatalysis Research with Structural Bioinformatics. In Applied Biocatalysis: From Fundamental Science to Industrial Applications; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2016; pp 71–100.

N. Schneider and A. Volkamer have written the book chapter and developed its concept. E. Nittinger has developed the classification of protein-protein interfaces and contributed to the manuscript. M. Rarey supervised this work.

D2 **Nittinger, E.**; Schneider, N.; Lange, G.; Rarey, M. Evidence of Water Molecules – A Statistical Evaluation of Water Molecules Based on Electron Density. J. Chem. Inf. Model. 2015, 55 (4): 771–783.

E. Nittinger has written the manuscript, developed and implemented the EDIA measurement. N. Schneider and G. Lange assisted during its design. G.Lange advised during the data set assembly concerning the used filter criteria. M. Rarey supervised this work.

D3 Meyder, A.; **Nittinger, E.**; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. J. Chem. Inf. Model. 2017, 57 (10): 2437-2447.

A. Meyder has written the manuscript, developed and conducted all the experiments. E. Nittinger has contributed to the EDIA$_m$ concept. G. Lange and R. Klein assisted the EDIA$_m$ design. M. Rarey supervised this work.

D4 Inhester, T.; **Nittinger, E.**; Sommer, K.; Schmidt, P.; Bietz, S.; Rarey, M. *NAOMI*nova: Interactive Geometric Analysis of Noncovalent Interactions in Macromolecular Structures. J. Chem. Inf. Model. 2017, 57 (9): 2132-2142.

T. Inhester and E. Nittinger have written the manuscript, developed and implemented *NAOMI*nova. K. Sommer helped compiling the use cases for the application of *NAOMI*nova. P. Schmidt contributed to the development of *NAOMI*nova throughout his master thesis. S. Bietz implemented the SMARTS pattern derivation used within *NAOMI*nova. M. Rarey supervised this work.

D5 **Nittinger, E.**; Inhester, T.; Bietz, S.; Meyder, A.; Schomburg, K.; Lange, G.; Klein, R.; Rarey, M. Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein-Ligand Interfaces. J. Med. Chem. 2017, 60 (10): 4245-4257.

E. Nittinger and T. Inhester have written the manuscript, developed and implemented *NAOMI*nova, which was used for the analysis of hydrogen bond geometries. E. Nittinger, T. Inhester and K. T. Schomburg designed the study. S. Bietz implemented the matching algorithm used in *NAOMI*nova. A. Meyder developed the application of EDIA for multiple atoms, which was integrated in *NAOMI*nova. R. Klein contributed to the area normalization calculations. G. Lange advised the data assembly. M. Rarey supervised this work.

D6 Fährrolfes, R.; Bietz, S.; Meyder, A.; **Nittinger, E.**; Otto, T.; Flachsenberg, F.; Volkamer, A.; Rarey, M. Proteins*Plus*: a web portal for structure analysis of macromolecules. Nucleic Acids Res. 2017, 45 (W1): W337-W343.

R. Fährrolfes and S. Bietz have written the manuscript. R. Fährrolfes has implemented the ProteinPlus server. S. Bietz developed SIENA and revised Protoss. A. Meyder and E. Nittinger developed the EDIA measurement for automatic structure validation with the experimental electron density data. E. Nittinger developed the hyPPI server for categorization of protein-protein interfaces. T. Otto refined PoseView for a 2D visualization of protein-ligand interactions. A. Volkamer developed the DoGSite scorer for the detection of druggable pockets in protein structures, which was further evolved by F. Flachsenberg. M. Rarey supervised this work.

D7 Bietz, S.; Inhester, T.; Lauck, F.; Sommer, K.; von Behren, M. M.; Fährrolfes, R.; Flachsenberg, F.; Meyder, A.; **Nittinger, E.**; Otto, T.; Hilbig, M.; Schomburg, K. T.; Volkamer, A.; Rarey, M. From cheminformatics to structure-based design: Web services and desktop applications based on the NAOMI library. Journal of Biotechnology. J. Biotechnol. 2017, 261: 207-214.

S. Bietz has written the manuscript with the assistance of all authors. S. Bietz developed SIENA and ASCONA and revised Protoss. T. Inhester developed PELIKAN for the analysis of spatial interaction patterns in protein structures. F. Lauck developed FSees for fragment space enumeration. K. Sommer developed UNICON for file format conversion. M. M. von Behren developed mRAISE for ligand-based virtual screening. R. Fährrolfes implemented the ProteinPlus server. A. Meyder and E. Nittinger developed the EDIA measurement for automatic structure validation with the experimental electron density data. E. Nittinger developed the hyPPI server for categorization of protein-protein interfaces. T. Otto refined PoseView for a 2D visualization of protein-ligand interactions. M. Hilbig developed MONA for the processing of data sets. K. T. Schomburg developed iRAISE for inverse virtual screening. K. T. Schomburg and S. Bietz finalized the SMARTSeditor. A. Volkamer developed the DoGSite scorer for the detection of druggable pockets in protein structures, which was further evolved by F. Flachsenberg. M. Rarey supervised this work.

D8 Schomburg, K. T.; **Nittinger, E.**; Meyder, A.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M. Prediction of protein mutation effects based on dehydration and hydrogen bonding – A large-scale study. Proteins Struct. Funct. Bioinforma. 2017, 85 (8): 1550-1566.

K. Schomburg has written the manuscript, developed and conducted all the experiments, and contributed to the concept of HYDE$_{protein}$. E. Nittinger has contributed to the manuscript, assembled the data set for the energy prediction experiment, and contributed to the design of the experiments. A. Meyder has contributed to the manuscript and implemented the HYDE optimizer. S. Bietz has contributed to the manuscript and helped implementing the HYDE mutation strategy. N. Schneider has implemented the prototype of the HYDE$_{protein}$ scoring method and initiated together with G. Lange, R. Klein, and M. Rarey, the concept of HYDE$_{protein}$. M. Rarey has supervised the project.

E1 von Behren, M. M.; Bietz, S.; **Nittinger, E.**; Rarey, M. mRAISE: An Alternative Algorithmic Approach to Ligand-Based Virtual Screening. Journal of Computer-Aided Molecular Design. Springer International Publishing August 26, 2016, pp 1–12.

M. M. von Behren has written the manuscript, developed and conducted all experiments. S.

Bietz and M. M. von Behren generated the alignment data set used for the evaluation of ligand superpositioning. S. Bietz integrated the necessary filter functionality into SIENA and developed the clustering approach for the identification of redundant ensembles. E. Nittinger contributed by supporting M. M. von Behren with the selection of high-resolution data for the validation of molecular alignments. M. Rarey supervised this work.

P1 **Nittinger, E.**; Flachsenberg, F.; Bietz, S.; Lange, G.; Rarey, M. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. J. Chem. Inf. Model. 2018, *Accepted*.

E. Nittinger has written the manuscript, developed and conducted the evaluation strategy. F. Flachsenberg has developed and implemented the optimization strategy, which was applied for the optimization of placed water molecule positions. G. Lange advised during the development of the water placement strategy. M. Rarey supervised this work.

P2 **Nittinger, E.**; Gibbons, P.; Eigenbrot, C.; Davies, D. R.; Maurer, B.; Yu, C. L.; Kiefer, J. R.; Kuglstatter, A.; Murray, J.; Ortwine, D. F.; Tang, Y.; Tsui, V. Water Molecules in Protein-Ligand Interfaces. Evaluation of Software Tools and SAR Comparison. J. Comput. Aided. Mol. Des., *Submitted for publication*.

E. Nittinger hast written the manuscript, developed and conducted the evaluation strategy. C. Eigenbrot, D. R. Davies, B. Maurer, C. L. Yu, J. R. Kiefer, A. Kugelstatter, J. Murray, and Y. Tang have resolved the BRD and BTK crystal structures released with this analysis. V. Tsui has supervised the project. D. F. Ortwine and P. Gibbons have contributed to the manuscript and have supervised the project.

A.2 Conference Contributions

This section lists the author's oral presentations and posters presented at national and international conferences.

Oral Presentations

T1 **Nittinger, E.**; Gibbons, P.; Tsui, V.; Rarey, M.; Ortwine, D. F. Conserved H2O in Protein-Ligand Complexes: An Evaluation of Water Prediction Tools. American Chemical Society (ACS) National Meeting. San Francisco, USA 2017.

T2 **Nittinger, E.**; Inhester, T.; Lange, G.; Klein, R.; Rarey, M. Hydrogen Bond Interaction Geometries in Proteins: From Big Statistics to Single Cases. American Chemical Society (ACS) National Meeting. San Francisco, USA 2017.

T3 **Nittinger, E.**; Inhester, T.; Lange, G.; Klein, R.; Rarey, M. Hydrogen Bond Interaction Geometries in Proteins: A Large-Scale Statistical Study. German Conference on Chemoinformatics (GCC). Fulda, Germany 2016.

T4 **Nittinger, E.**; Schomburg, K. T.; Lange, G.; Rarey, M. Protein-Ligand Interaction Preferences: An Evaluation of Actual vs. Possible Hydrogen Bonds. MGMS Young Modellers' Forum. London, Great Britain 2015.

T5 **Nittinger, E.**; Schomburg, K. T.; Lange, G.; Rarey, M. Protein and Ligand Interaction Preferences: A Large Scale Study. Gordon Research Seminar – Computer Aided Drug Design (GRS CADD). Mount Snow, USA 2015.

T6 **Nittinger, E.**; Schneider, N.; Lange, G.; Rarey, M. EDIA – A New Estimate of Electron Density of Individual Atoms for Validating Water Molecules. 29th Molecular Modelling Workshop. Erlangen, Germany 2015.

T7 **Nittinger, E.**; Schneider, N.; Lange, G.; Rarey, M. The Motility of Water Molecules – A Statistical Evaluation of Water Molecules Based on Electron Density. International Conference on Chemical Structures (ICCS). Noordwijkerhout, Netherlands 2014.

Poster Presentations

P1 **Nittinger, E.**; Lange, G.; Rarey, M. HYDE – Modeling Water Molecules in Protein Complexes. Gordon Research Conference – Computer Aided Drug Design (GRC CADD). Mount Snow, USA 2017.

P2 **Nittinger, E.**; Inhester, T.; Rarey, M. *NAOMI*nova: Interactive Analysis of Non-Covalent Interactions in Protein Complexes. Gordon Research Seminar – Computer Aided Drug Design (GRS CADD). Mount Snow, USA 2017.

P3 **Nittinger, E.**; Gibbons, P.; Tsui, V.; Ortwine, D.; Rarey, M. „Water, Water, Everywhere...“: Modeling Water Molecules in Protein-Ligand Complexes. American Chemical Society (ACS) National Meeting. San Francisco, USA 2017.

P4 **Nittinger, E.**; Gibbons, P.; Tsui, V.; Ortwine, D. "Water, Water, Everywhere..." An Evaluation of Water Prediction Tools. German Conference on Chemoinformatics (GCC). 2016.

P5 **Nittinger, E.**; Schomburg, K.; Lange, G.; Rarey, M. Protein and Ligand Interaction Preferences: A Large Scale Study. Gordon Research Conference – Computer Aided Drug Design (GRC CADD). Mount Snow, USA 2015.

P6 **Vennmann, E.**; Schneider, N.; Lange, G.; Rarey, M. Elucidating Protein-Protein Interactions Using the HYDE Scoring Function. German Conference on Chemoinformatics (GCC). Fulda, Germany 2013.

P7 **Vennmann, E.**; Schneider, N.; Lange, G.; Rarey, M. Protein-Protein Interactions: Interface Hotspots and Classification. 7th Vienna Summer School on Drug Design. Vienna, Austria 2013.

B
Methodical Details

In this chapter, additional details about the methods not included in the publications of this dissertation are described.

B.1 HYDE Scoring Function

The HYDE scoring function was originally developed to score protein-ligand complexes.[15–19] Due to its generic concept with no training on experimental binding affinity data, it can also be applied to score single amino acids, whole proteins, or protein-protein complexes. In this section, the alterations implemented in the course of this dissertation will be described. For further information on the basic concept of HYDE please refer to the dissertation of N. Schneider.[17]

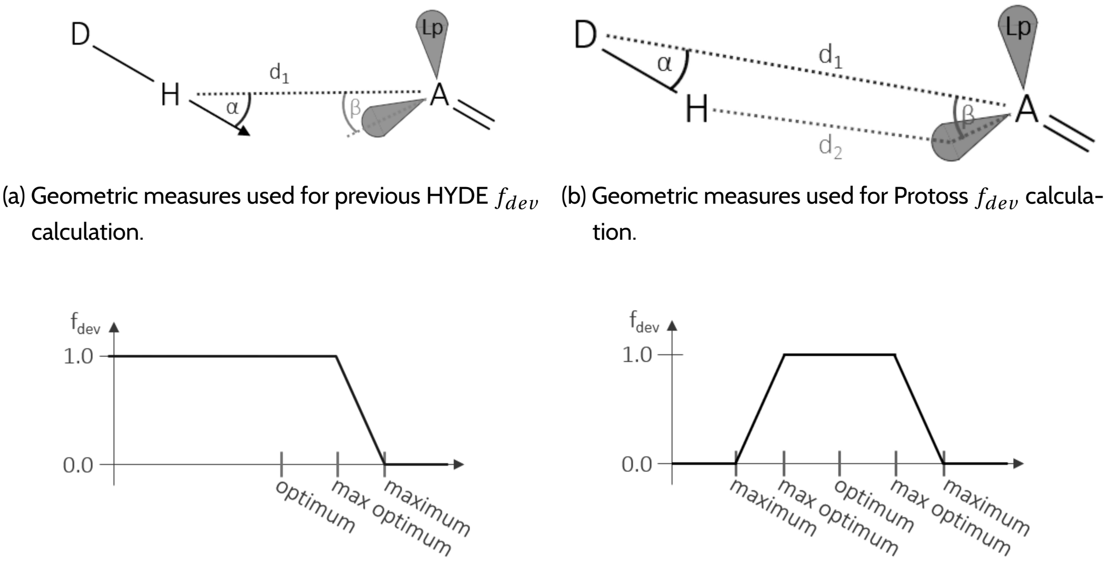
B.1.1 Details of the HYDE Implementation

For differentiation, the previous HYDE implementation will be referred to as HYDE$_{2012}$.

Assimilation of HYDE and Protoss

Protoss is used by default to optimize the hydrogen bond network before HYDE scoring. Protoss and HYDE$_{2012}$ used different scoring schemes (Figures B.1a and B.1b), which let to different qualitative assessments between Protoss and HYDE$_{2012}$. The angles α and β indicated in Figure B.1b were measured in our large-scale analysis of hydrogen bond interaction patterns (D5). Therefore, the

HYDE scoring scheme was changed to the Protoss scoring scheme including an adaption of the ideal and maximum angle values. HYDE$_{2012}$ applied an asymmetric scoring of hydrogen bond distances (Figure B.1c), which let to no penalty for clashing atoms. The calculation of f_{dev} for distances was changed to a symmetric scoring. Based on these alterations, HYDE and Protoss now use the same scoring scheme and qualitatively achieve the same geometric measures.



(a) Geometric measures used for previous HYDE f_{dev} calculation.

(b) Geometric measures used for Protoss f_{dev} calculation.

(c) Asymmetric scoring of hydrogen bond distances.

(d) Symmetric scoring for hydrogen bond distances and angles.

Figure B.1: Adaption of geometric measures to align HYDE and Protoss.

Adaption of $\log$ P Parameter

The p$\log$P parameters for polar and apolar atoms can be derived from the HYDE concept. Based on the thermodynamic cycle of water, the total energy for breaking all four hydrogen bonds, i.e. from ice to vapor, is 54.18 kJ mol^{-1} (Figure B.2).

This concludes that breaking one hydrogen bond function costs 13.55 kJ mol^{-1} or expressed favorably, forming a hydrogen bond leads to an energy gain of ΔG_{sat} = -13.55 kJ mol^{-1}. The fraction of unsaturated hydrogen bonds in water, as explained in Chapter 4.2, is the reason for the stabilizing energetic contribution upon formation of a hydrogen bond. According to HYDE, the energy contribution of a polar atoms can be calculated as follows:

$$\Delta G_{\text{saturation}} = \frac{2.3 \cdot RT}{f_{sat}} \cdot p\log P^{i} \cdot \sum_{\text{HB j}} w^{j} \cdot f_{dev}^{j} \tag{B.1}$$

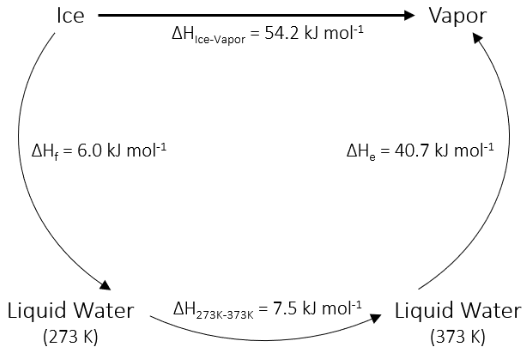

Ice
Vapor
ΔH$_{Ice-Vapor}$ = 54.2 kJ mol^{-1}
ΔH$_f$ = 6.0 kJ mol^{-1}
ΔH$_e$ = 40.7 kJ mol^{-1}
Liquid Water
(273 K)
ΔH$_{273K-373K}$ = 7.5 kJ mol^{-1}
Liquid Water
(373 K)

Figure B.2: Thermodynamic cycle of water; ΔH_f = enthalpy of fusion, ΔH_f = enthalpy of heating water from 273 K to 373 K, ΔH_e = enthalpy of evaporation.

Assuming an ideal interaction geometry (f_{dev} = 1) and only one hydrogen bond function (w^j = 1), the equation can be solved for plog P accordingly:

$$
\begin{aligned}
p\log P &= \frac{\Delta G_{saturation} \cdot f_{sat}}{2.3 \cdot R \cdot T} \\
&= \frac{-13.58 kJmol^{-1} \cdot 0.85}{2.3 \cdot 8.314 Jmol^{-1}K^{-1} \cdot 298K} \\
&= -2.032
\end{aligned}
\tag{B.2}
$$

The introduction of an apolar functional group disrupts the hydrogen bond network of pure water. Based on the HYDE concept it is assumed that an idealized apolar group leads to one unsatisfied functional group of a water molecule while the remaining three hydrogen bond functions are satisfied. Due to a temperature dependent factor difference between the entropy and enthalpy, an overall unfavorable free energy results for the introduction of apolar groups into water. To avoid unsatisfied hydrogen bond functions, apolar moieties aggregate. Thus the removal of an unsatisfied hydrogen bond function, i.e. the dehydration of an apolar group, results in a favorable free energy contribution, known as the hydrophobic effect. The free energy contribution of an apolar atom is $\Delta G_{dehyd}^{apolar}$ = -2.7 kJ mol^{-1} according to the HYDE theory. Since the HYDE theory assumes that exactly one hydrogen bond function is covered by introduction of an idealized apolar moiety, the energy gain results from the removal of an unsatisfied hydrogen bond function. To approximate the surface area of one hydrogen bond function, the covered molecular surface area of water molecules of ten protein structures was calculated and related to their number of formed hydrogen bonds. Water molecules participating in one hydrogen bond have a minimum of 6.81 Å^2 covered surface area. This leads to 25.13 Å^2 covered solvent accessible surface area for one hydrogen bond function of a water molecule. The formula for calculating the free energy of dehydration of apolar atoms is as follows and can be solved for plog P accordingly:

$$\Delta G_{dehyd}^{apolar} = -2.3 \cdot R \cdot T \cdot p\log P \cdot \Delta acc \tag{B.3}$$

$$\begin{aligned} p\log P &= \frac{\Delta G_{dehyd}}{-2.3 \cdot R \cdot T \cdot \Delta acc} \\ &= \frac{-2.7kJmol^{-1}}{-2.3 \cdot 8.314Jmol^{-1}K^{-1} \cdot 298K \cdot 25.13\text{Å}^2} \\ &= 0.0188 \end{aligned} \tag{B.4}$$

This leads to an energy gain of -107 J mol^{-1} Å^{-2}, which agrees well with experimentally obtained values ranging from -67 J mol^{-1} Å^{-2} (with no temperature specified),[89] -125 – -138 J mol^{-1} Å^{-2}, to -119 – -149 J mol^{-1} Å^{-2} (at room temperature).[90,91]

The plogP values based on the HYDE theory were used for the HYDE calculations.

B.2 NAOMI Interaction Framework

The NAOMI interaction framework is used for generating and scoring hydrogen bond interactions. In this section, the new implementations for discretization interaction surfaces will be described. For details on the interaction framework itself please refer to the dissertation of S. Bietz.[310]

B.2.1 Interaction Surface Discretization

Interaction directions and their assigned geometries are defined using chemical type (ChemType) and geometry type (GeomType) definitions. The ChemType defines the interacting atom, e.g. a nitrogen acceptor, a carbonyl oxygen etc.. The GeomType specified the geometry of the interaction surface, e.g. cone, spherical rectangle, or capped cone. For the combination of ChemType and GeomType interaction deviations are available. Those interaction deviations consist of an optimum, the maximal optimum deviation (maxOpt) up to which an interaction is ideal, and the maximum until which the score contribution is decreased to zero.

Based on the assigned interaction surfaces a discretization with concentric circles is calculated (Figure B.3).

1. First Rotation

   a) Define angle α with dot distance d_D

   $\alpha = 360°/round\left(\frac{U_1}{d_D}\right)$

   $U_1 = 2\pi \cdot r_1$

   b) Define number of concentric circles *#CC*

   Cone: $\#CC = maxDev/\alpha$

   Rectangle: $\#CC = max(maxDev_{in}, maxDev_{out})/\alpha$

   c) Define rotation axis $\overrightarrow{r_1}$ for rotation

   $\overrightarrow{r_1} = \vec{v} \times \overrightarrow{ia}$

   d) Calculate first position on concentric circle

   - Check if position is within rectangle range

2. Second rotation (on concentric circle with index i)

   a) Define angle β with number of points on concentric circle p_{CC}

   $\beta = 360°/p_{CC}$

   $p_{CC} = round\left(\frac{U_2}{d_D}\right)$

   $U_2 = 2\pi \cdot r_2$

   $r_2 = \sin(\alpha \cdot i) \cdot r_1$

   b) Define rotation axis $\overrightarrow{r_2}$

   $\overrightarrow{r_2} = \overrightarrow{ia}$

   c) For rectangle geometry:
   check if position is within rectangle range

3. Repeat 1 and 2 until i = #CC

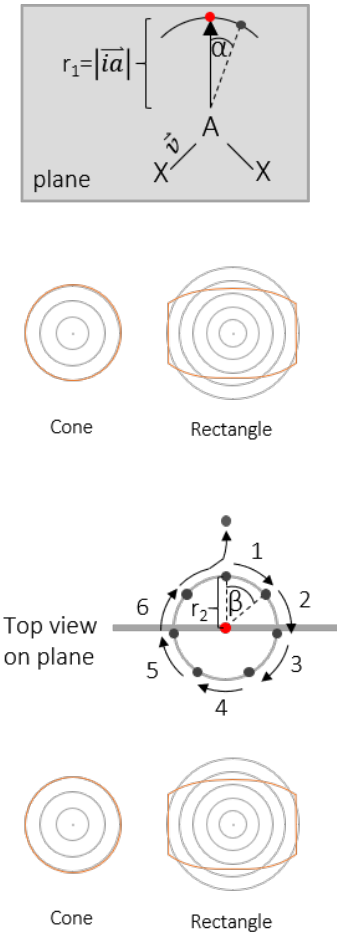


Figure B.3: Discretization of interaction surfaces.

C
Software Architecture and Application

In this chapter, the developed software architecture and the application programs based on them are presented. All components were based on the previously developed NAOMI framework.[267]

An overview of the dependencies of software applications and the main developed, altered, and used software libraries is given in Figure C.1. PROTOSS is a previously developed software library used for all presented applications and necessary for the developed free space identification and water placement procedure. The INTERACTIONS and HYDE software library are existing software libraries whose functionalities have been altered or extended. The software libraries CRYSTALGEOMETRY as well as WATERPREDICTION were implemented for the interpretation of electron density maps, the EDIA calculation as well as water placement procedure.

The aim of this chapter is to provide insight into the decisions made during the development as well as their internal dependencies. The chapter is organized in two sections: (1) The developed internal libraries with their concepts, classes, and dependencies. (2) An overview of the new and extended application programs.

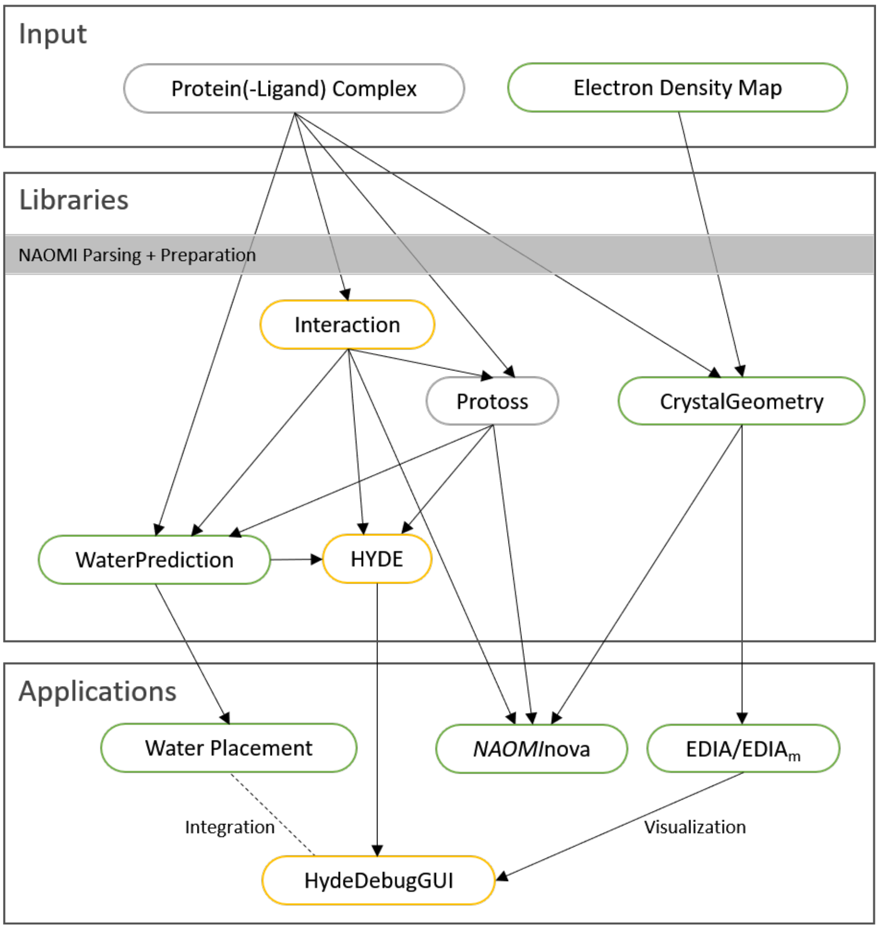


Figure C.1: General structure of software libraries and dependencies of the NAOMI library; green = new development; yellow = altered or extended; gray = used.

C.1 Software Libraries

Crystal Geometry Library

The CRYSTAL GEOMETRY library contains functionalities for the analysis and interpretation of crystal structure information. The library provides an interface for three main functionalities:

1. The storage of electron density values parsed from electron density maps.

2. The calculation of EDIA and EDIA$_m$ based on electron density values.

3. The generation of crystal symmetry contacts based on crystal contact information from PDB files.

The main class ElectronDensityData is used for storing information of electron density maps, such as the electron density grid, the origin of grid, the number of data points in each direction x, y, and z, or the angles between x, y, and z axis. The main component is the 3D grid with its assigned electron density values. Based on the stored density information in ElectronDensityData, the class ElectronDensityScorer calculates EDIA and EDIA$_m$ scores.

The generation of crystal symmetry complexes depends on the *CRYST1* information given in the header section of PDB files:

```
CRYST1 30.800 45.000 40.300 90.00 90.00 90.00 P 21 21 2
```

The first three number are the unit cell lengths (a, b, and c). The following three numbers are the angles α, β, and γ, with α between directions b and c, β between directions a and c, and γ between directions a and b. The remaining information indicates the space group (P 21 21 2). For all space groups, the library contains necessary translation vectors and transformation matrices in the SymmetryGeneratorList. Based on a definable distance threshold, the crystal contacts surrounding the protein complex from the PDB file are generated.

HYDE Library

The HYDE library provides the interface for scoring protein-ligand interactions as well as HYDE$_{protein}$ for the estimation of protein-protein affinities as well as the protein itself.

The main components of the HYDE library are:

- HydeComponentBuilder to build the two textscHydeComponents necessary for scoring: one represents the active site, the other one the molecule to be scored.

- HydeScorerContextBuilder to initialize the scoring context, i.e. the identification of close atoms, the generation of interactions, as well as matching interactions.

- HydeScorer to calculate the HYDE score based on the HydeHydrophilicScorePolicy, which determines the scoring of polar atoms, and the HydeHydrophobicScorePolicy, which determines the scoring of apolar atoms.

In order to generate the two HydeComponents two active sites are necessary. The first active site is used as the area in which the HYDE score is calculated. The second active site needs to be bigger than the first one and is necessary for the correct calculation of the change in surface area, which is used for the calculation of apolar atom contributions.

The calculated HYDE scores of the HydeComponent representing the active site, can be mapped onto the second HydeComponent using the MappedHydeScore class. This way, the atom scores of the surrounding are mapped on the closest atom of the molecule.

Further classes, HydeProteinInterfaceScore and HydeProteinScore, generate HYDE scores for protein-protein interfaces and the protein itself, respectively.

Interactions Library

The INTERACTIONS library contains all functionalities around interactions, from calculating interaction directions, discretization of interaction surfaces, to scoring interactions. In this section, only the altered and new implementations will be described. For detailed information about the remaining functionality please refer to the dissertation of S. Bietz.[310]

Based on the large scale analysis of hydrogen bond interactions in protein-ligand active sites (D5) additional chemical types (CHEMTYPES) to the already existing NAOMI CHEMTYPES were defined:

- IASTATCARBONYL: ketones and primary amide oxygen atoms.
- IASTATCARBOXYL: carboxylic acid oxygen atoms.
- IASTATAMIDECARBONYL: secondary and tertiary amide oxygen atoms.
- IASTATESTER: sp^2-hybridized ester oxygen atom.
- IASTATETHER: ether oxygen atom as well as ester sp^3 hybridized oxygen atom.
- IASTATACCWOPLANE: an acceptor, where no plane can be defined.

The main alterations concern sp^2-hybridized oxygen acceptors that showed distinct differences in their opening angle, i.e. the angle between the electron lone pairs. In addition to the CHEMTYPES, angle deviation were defined (STATISTICINTERACTIONDEVIATIONS).

Based on the new CHEMTYPES in combination with geometry types (GEOMTYPES: cone, spherical rectangle, or capped cone) interaction surfaces are defined. These interaction geometries are discretized by the class INTERACTIONSURFACELIST. This class is implemented as a singleton, thus its instance is only created ones. Every used combination of CHEMTYPE and GEOMTYPE is only calculated once and stored for further use. Herein, the interaction direction is translated onto the x-axis and the in-plane angle (first rotation direction) is rotated onto the y-axis. All discretized surface points are transformed respectively. The individual points of the discretized interaction surface are stored as pairs POINTSCOREPAIR, containing the coordinate of the point and its assigned geometric score. The geometric score is based on angles and distances defining a hydrogen bond interaction (for an example see B.1b). The use of a template allows a user defined INTERACTIONSCORER, which defines the distances and angles as well as their combination to a final geometric score. All points of the discretized interaction surface are stored in a vector POINTSCOREVECTOR, sorted by the geometric score of the POINTSCOREPAIRS. This allows a consistent storage of discretized interaction surfaces and the transformation back onto other interaction directions. After transformation onto the current interaction direction, the availability of the discretized surface points needs to be calculated, i.e. the overlap with surrounding protein or ligand atoms.

Water Prediction Library

The WATERPREDICTION library provides functionalities for implicit as well as explicit water molecules. In case of implicit water molecules, free space needs to be identified correctly. Based on discretized interaction surfaces, the free space is identified. Using the identification of free space, explicit water molecules can be placed.

The PotentialWaterScorer class provides the interface to the InteractionsLibrary and its discretization of interaction surfaces used for implicit water molecules. Three main functionalities are provided:

F1 Returns all transformed discretized points with their corresponding geometry scores for a given interaction direction.

F2 Returns the best available position for an implicit water molecule on the discretized interaction surface.

F3 Returns the best geometric score for an implicit water molecule.

The WaterPlacement class needs InteractionDeviations and an InteractionScorer to place water molecules. Either for a specified active site or for the whole protein complex water molecules can be placed. Free interaction directions of the active site or protein are identified and using the function F1 of the PotentialWaterScorer, possible water positions are assigned. This information is subsequently used for clustering and generation of explicit water molecules.

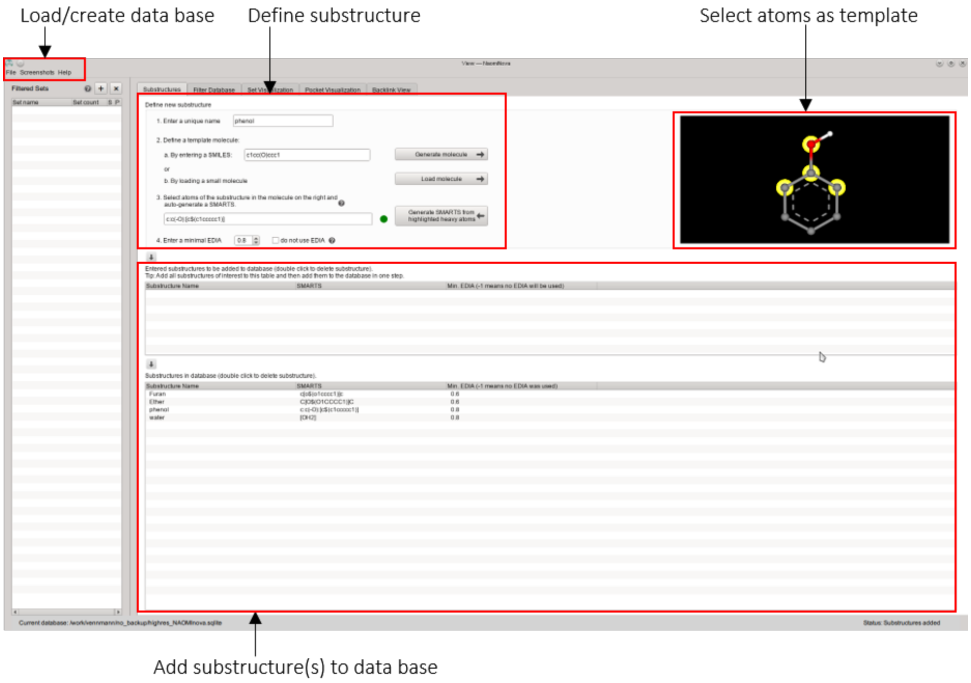
C.2 Software Application

This chapter explains the basic program options for *NAOMI*nova, HydeDebugGUI and PPI.

C.2.1 *NAOMI*nova

*NAOMI*nova was developed for large scale analysis of user-defined substructures. The developed graphical user interface (GUI) allows diverse opportunities for the generation of substructures of interest as well as their further structural analysis.

First, a user-defined set of protein structures has to be selected for the generation of the SQLite data base (Figure C.2). Once a data base has been created, it can be loaded an re-used. Next, a substructure of interest has to be defined by the user (Figure C.2). Herein, two different possibilities exist: Either a small molecule is loaded from a sdf or mol2 file or a SMARTS can be supplied. Both ways, the molecule is generated and shown in a 3D view, in which the desired atoms can be selected. Selected atoms can further be specified using the full range of the SMARTS language. This allows the definition of a surrounding, without the need to specify the exact geometry. In addition to the substructure definition, a unique name and optional a minimal EDIA value are chosen. Using the minimal EDIA criterion allows a qualitative evaluation at a later stage. Once one or multiple substructures have been defined, they must be added to the data base. The previously selected atoms will be used for superimposing matching structures from the data base.



Define new substructure

1. Enter a unique name: phenol

2. Define a template molecule:

a. By entering a SMILES: c1cc(O)ccc1

or

b. By loading a small molecule

3. Select atoms of the substructure in the molecule on the right and auto-generate a SMARTS.

c:c(-O):[c$(c1ccccc1)]

4. Enter a minimal EDIA: 0.8 □ do not use EDIA

Entered substructures to be added to database (double click to delete substructure).
Tip: Add all substructures of interest to this table and then add them to the database in one step.

Substructure Name	SMARTS	Min. EDIA (-1 means no EDIA will be used)

Substructures in database (double click to delete substructure).

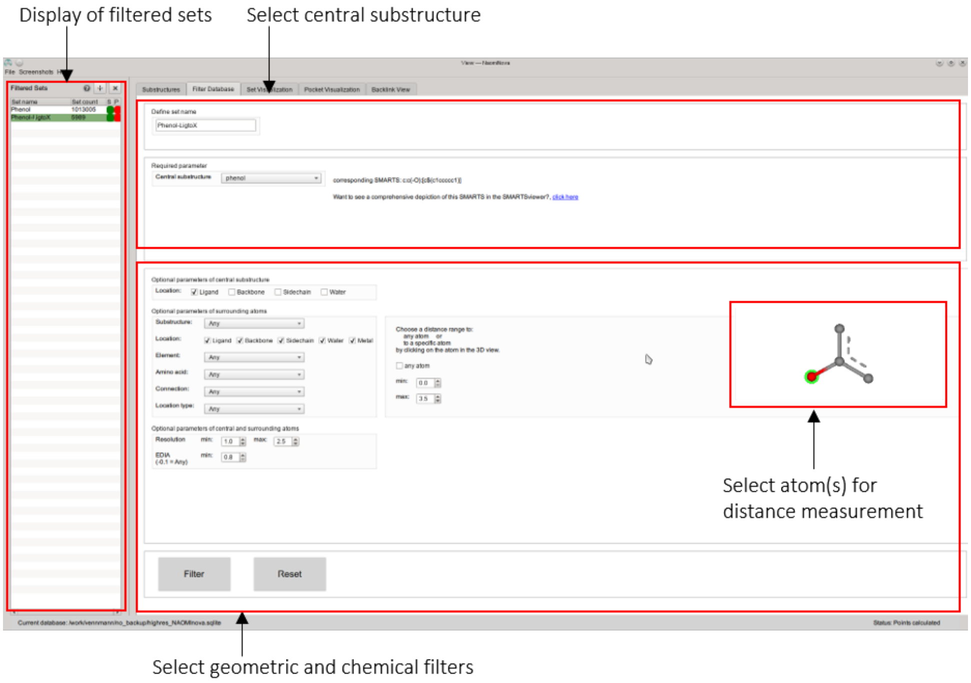
Substructure Name	SMARTS	Min. EDIA (-1 means no EDIA was used)
Furan	c[o$(o1cccc1)]c	0.6
Ether	C[O$(O1CCCC1)]C	0.6
phenol	c:c(-O):[c$(c1ccccc1)]	0.8
water	[OH2]	0.8

Current database: /work/vennmann/no_backup/highres_NAOMInova.sqlite

Status: Substructures added

Figure C.2: *NAOMI*nova substructure definition tab.

After adding substructures to the data base, they can be filtered for diverse chemical and geometric criteria (Figure C.3). The only mandatory parameter is the selection of a functional group as central substructure. All further parameters are optional and range from location, i.e. backbone, side chain, molecule, or water, to connection, i.e. inter or intra. Additionally, the resolution of the original PDB can be specified as well as the EDIA, if it was selected during substructure definition.


Display of filtered sets

Select central substructure

Filtered Sets

Set name	Set count
Phenol	1013006
Phenol-LigtoX	5989

Define set name: Phenol-LigtoX

Required parameter

Central substructure: phenol

corresponding SMARTS: c:c(-O):[c$(c1ccccc1)]

Want to see a comprehensive depiction of this SMARTS in the SMARTSviewer?, click here

Optional parameters of central substructure

Location: ☑ Ligand ☐ Backbone ☐ Sidechain ☐ Water

Optional parameters of surrounding atoms

Substructure: Any

Location: ☑ Ligand ☑ Backbone ☑ Sidechain ☑ Water ☑ Metal

Element: Any

Amino acid: Any

Connection: Any

Location type: Any

Choose a distance range to:
any atom or
to a specific atom
by clicking on the atom in the 3D view.

☐ any atom

min: 0.0

max: 3.5

Optional parameters of central and surrounding atoms

Resolution min: 1.0 max: 2.5

EDIA (-0.1 = Any) min: 0.8

Select atom(s) for distance measurement

Current database: /work/vennmann/no_backup/highres_NAOMInova.sqlite

Status: Points calculated

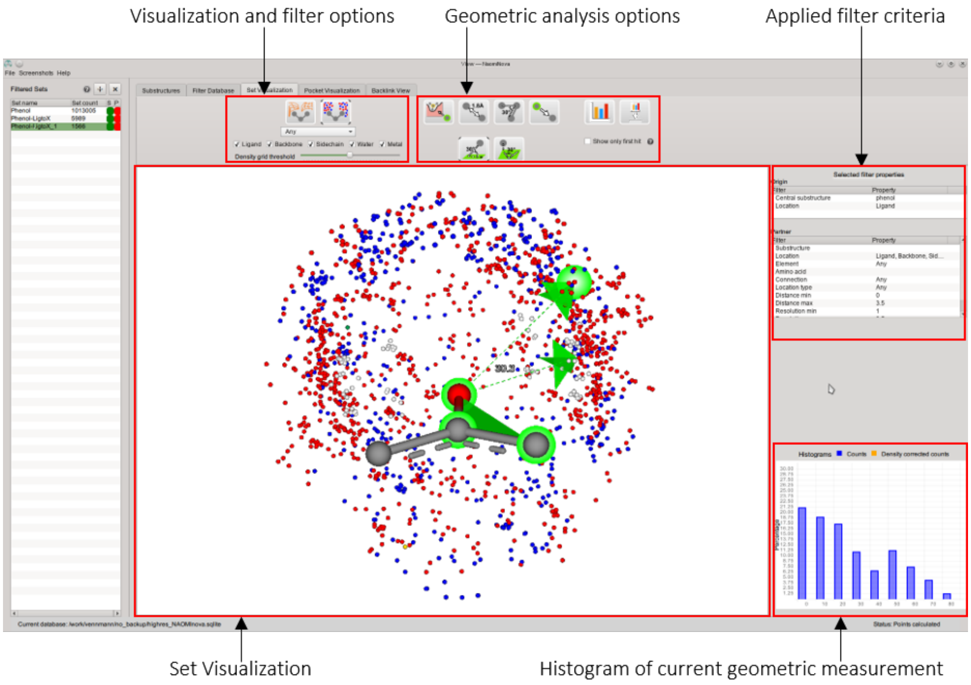
Select geometric and chemical filters

Figure C.3: *NAOMI*nova filter tab.

Once the filter process is completed, the sets are shown in the list view on the left hand side of the GUI (Figure C.4). Now, the user has to options: (1) the set can be analyzed for geometric and chemical criteria in the *Set View* (Figure C.4) or (2) a protein-ligand active site can be loaded and suitable filtered sets can be superimposed into the active site (Figure C.5).

The *Set View* contains diverse further filter criteria that can be applied to the already pre-filtered set as well as the geometric analysis and generation of volume-normalized histograms, i.e. for the evaluation of preferred positions of partner points around the central substructure.

The *Pocket View* contains the same additional filter criteria as the *Set View*. Based on the superimposed set and pocket, the user can identify regions, where partner points are preferred and those, where no partner points can be observed based on the structures in the data base.



Set name	Set count	S	P
Phenol	1013006		
Phenol-LigtoX	5989		
Phenol-LigtoX_1	1566		

☑ Ligand ☑ Backbone ☑ Sidechain ☑ Water ☑ Metal

Selected filter properties

Origin

Filter	Property
Central substructure	phenol
Location	Ligand

Partner

Filter	Property
Substructure	
Location	Ligand, Backbone, Sid...
Element	Any
Amino acid	
Connection	Any
Location type	Any
Distance min	0
Distance max	3.5
Resolution min	1

Figure C.4: *NAOMInova* set visualization tab.

Finally, every partner point that is either displayed in the *Set* or *Pocket View* is connected to is source. This means, that the protein structure of the partner point can be displayed for further analysis, so-called 'backlink' (Figure C.6). This allows the validation of the source and by this if the partner point is of interest for answering my questions.

Overall, *NAOMInova* a is versatile tools for the analysis of user-defined substructures of interest with a diverse spectrum of filter criteria that aid its geometric and chemical evaluation.

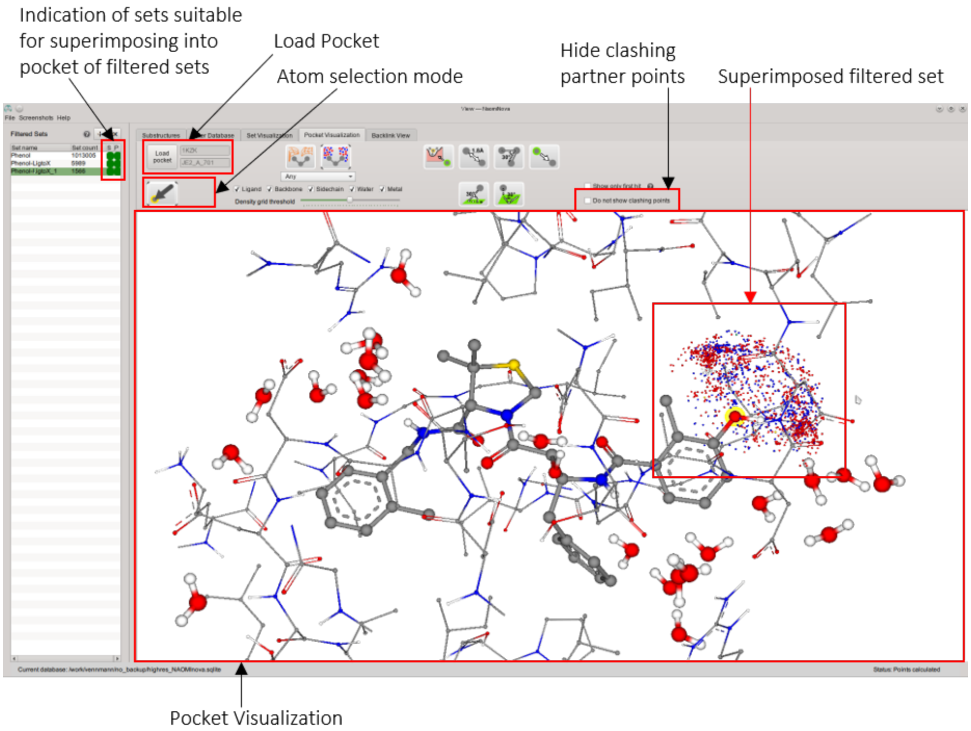

Indication of sets suitable for superimposing into pocket of filtered sets
Load Pocket
Atom selection mode
Hide clashing partner points
Superimposed filtered set
Pocket Visualization

Figure C.5: *NAOMI*nova pocket visualization tab.

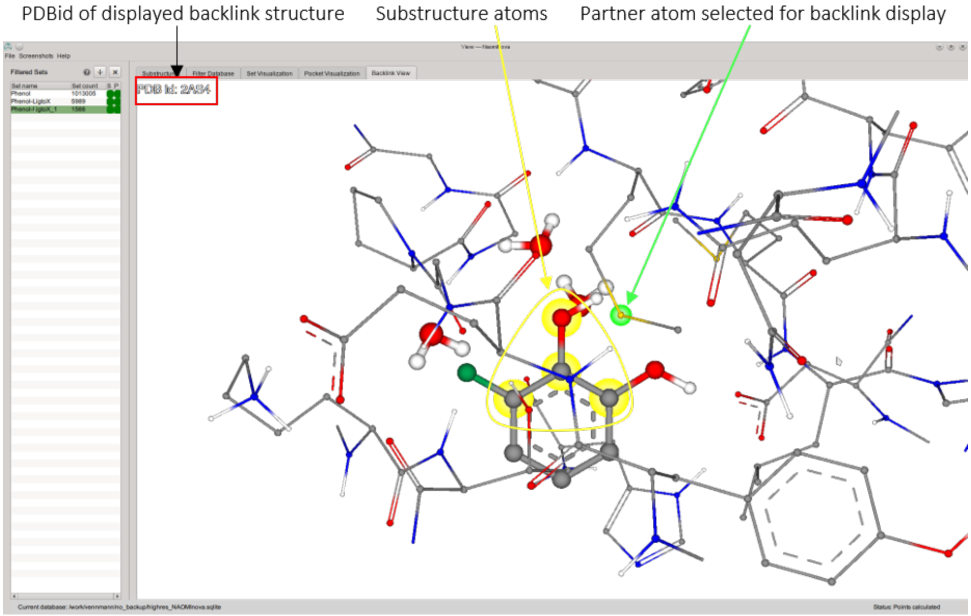

PDBid of displayed backlink structure
Substructure atoms
Partner atom selected for backlink display
PDB Id: 2A84

Figure C.6: *NAOMI*nova backlink visualization.

C.2.2 HYDE Debug GUI

The HYDE Debug GUI was developed to visualize the results of the scoring function HYDE. HYDE calculated results are atom-based and allow an easy-to-interpret visualization (Figure C.7).

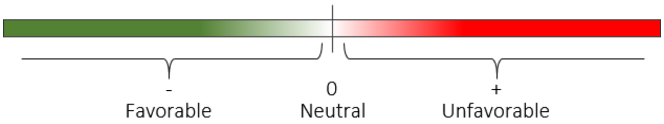
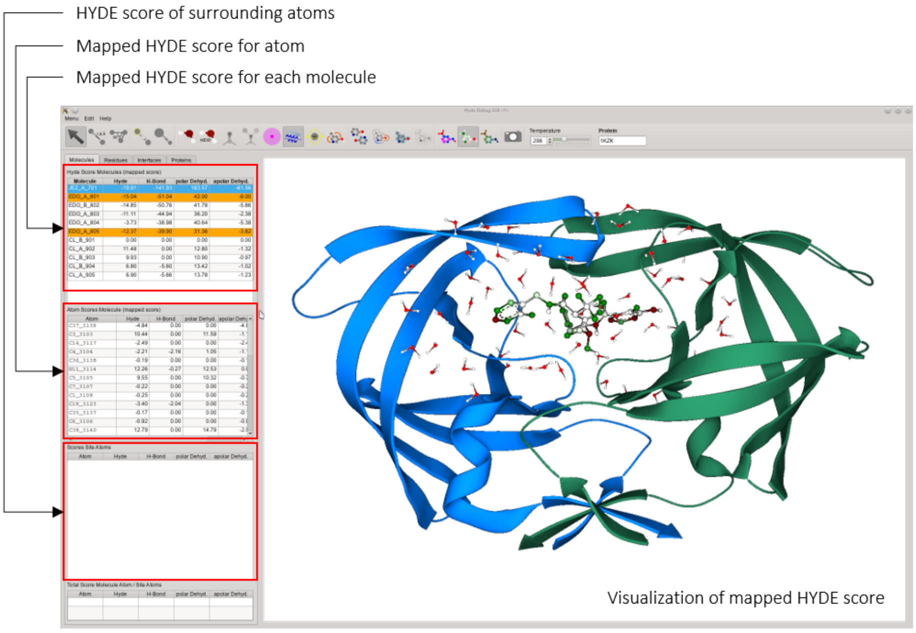


Figure C.7: HYDE coloring scheme.

Originally, HYDE was developed to score protein-ligand complexes (Figure C.8). Due to no training on experimental data such as small molecule binding affinities, HYDE can also be applied for scoring proteins (Figure C.9) – in whole, but also single amino acids, or individual atoms – or protein-protein interactions (Figure C.10).

First, a protein structure has to be loaded or fetched directly from the PDB. If the protein structure is fetched, the electron density data (2fo-fc and fo-fc map) will be retrieved automatically as well, if it is available. Otherwise the electron density data can be loaded separately.

The *Molecule View* displays the mapped HYDE score, i.e. the HYDE scores of the pocket atoms are mapped onto the closest ligand atom (Figure C.8). The predicted affinity can be analyzed and broken down to the individual atom contributions. For the protein-ligand affinity prediction, the user has the option to geometrically optimize the ligand with GeoHYDE.


HYDE score of surrounding atoms
Mapped HYDE score for atom
Mapped HYDE score for each molecule
Hyde Debug GUI <5>
Menu
Edit
Help
Temperature
298
Protein
1KZK
Molecules
Residues
Interfaces
Proteins
Visualization of mapped HYDE score

Hyde Score Molecules (mapped score)

Molecule	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.
JE2_A_701	-19.91	-141.93	183.57	-61.56
EDO_A_801	-15.04	-51.04	42.00	-6.00
EDO_B_802	-14.85	-50.76	41.78	-5.86
EDO_A_803	-11.11	-44.94	36.20	-2.38
EDO_A_804	-3.73	-38.98	40.64	-5.38
EDO_A_805	-12.37	-39.90	31.36	-3.82
CL_B_901	0.00	0.00	0.00	0.00
CL_A_902	11.48	0.00	12.80	-1.32
CL_B_903	9.93	0.00	10.90	-0.97
CL_B_904	6.80	-5.60	13.42	-1.02
CL_A_905	6.90	-5.66	13.78	-1.23

Atom Scores Molecule (mapped score)

Atom	Hyde	H-Bond	polar Dehyd.	apolar Dehy
C37_3139	-4.84	0.00	0.00	-4.
C1_3103	10.44	0.00	11.59	-1.
C14_3117	-2.49	0.00	0.00	-2.
C4_3104	-2.21	-2.16	1.05	-1.
C36_3138	-0.19	0.00	0.00	-0.
H11_3114	12.26	-0.27	12.53	0.
C5_3105	9.55	0.00	10.32	-0.
C7_3107	-0.22	0.00	0.00	-0.
C1_3108	-0.25	0.00	0.00	-0.
C19_3123	-3.40	-2.04	0.00	-1.
C35_3137	-0.17	0.00	0.00	-0.
C6_3106	-0.92	0.00	0.00	-0.
C38_3140	12.79	0.00	14.79	-2.

Scores Site Atoms

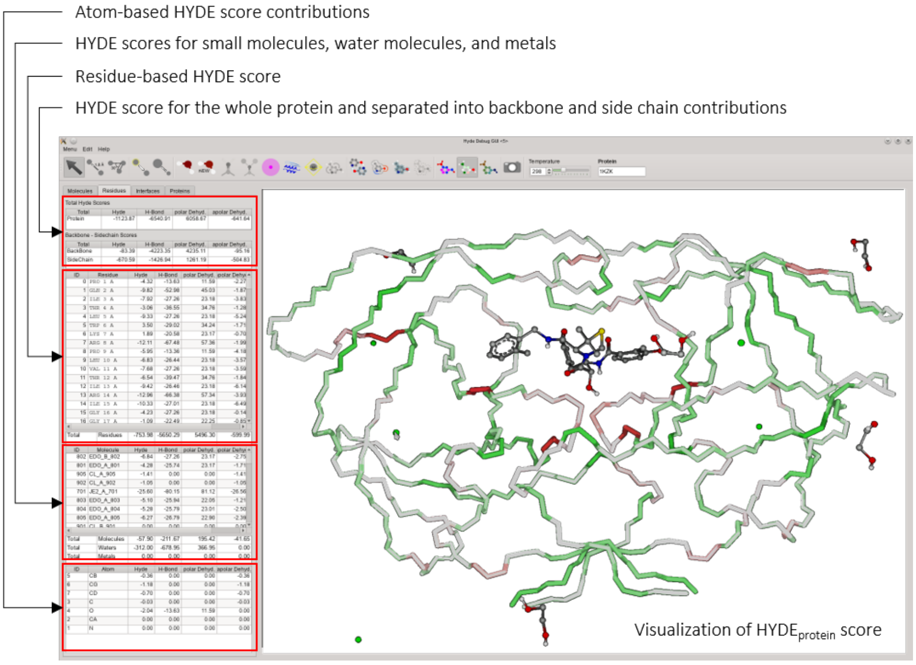
Atom	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.

Total Score Molecule Atom / Site Atoms

Atom	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.

Figure C.8: HYDE Debug GUI molecule tab.

HYDE$_{protein}$ scores can be analyzed in the *Protein View* (Figure C.9). These scores resemble the difference between the unfolded and folded protein. The contribution of the whole protein can be analyzed and split into contributions from backbone as well as side chains. Additionally, each individual amino acid can be analyzed as well as ligands, co-factors, water molecules or metals.


Atom-based HYDE score contributions
HYDE scores for small molecules, water molecules, and metals
Residue-based HYDE score
HYDE score for the whole protein and separated into backbone and side chain contributions
Hyde Debug GUI <5>
Menu
Edit
Help
NEW
Molecules
Residues
Interfaces
Proteins
Visualization of HYDE$_{protein}$ score

Temperature: 298

Protein: 1KZK

Total Hyde Scores

Total	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.
Protein	-1123.87	-6540.91	6058.67	-641.64

Backbone - Sidechain Scores

Total	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.
BackBone	-83.39	-4223.35	4235.11	-95.16
SideChain	-670.59	-1426.94	1261.19	-504.83

ID	Residue	Hyde	H-Bond	polar Dehyd.	polar Dehy
0	PRO 1 A	-4.32	-13.63	11.59	-2.27
1	GLN 2 A	-9.82	-52.98	45.03	-1.87
2	ILE 3 A	-7.92	-27.26	23.18	-3.83
3	THR 4 A	-3.06	-36.55	34.76	-1.28
4	LEU 5 A	-9.33	-27.26	23.18	-5.24
5	TRP 6 A	3.50	-29.02	34.24	-1.71
6	LYS 7 A	1.89	-20.58	23.17	-0.70
7	ARG 8 A	-12.11	-67.48	57.36	-1.99
8	PRO 9 A	-5.95	-13.36	11.59	-4.18
9	LEU 10 A	-6.83	-26.44	23.18	-3.57
10	VAL 11 A	-7.68	-27.26	23.18	-3.59
11	THR 12 A	-6.54	-39.47	34.76	-1.84
12	ILE 13 A	-9.42	-26.46	23.18	-6.14
13	ARG 14 A	-12.96	-66.38	57.34	-3.93
14	ILE 15 A	-10.33	-27.01	23.18	-6.49
15	GLY 16 A	-4.23	-27.26	23.18	-0.14
16	GLY 17 A	-1.09	-22.49	22.25	-0.85
Total	Residues	-753.98	-5650.29	5496.30	-599.99

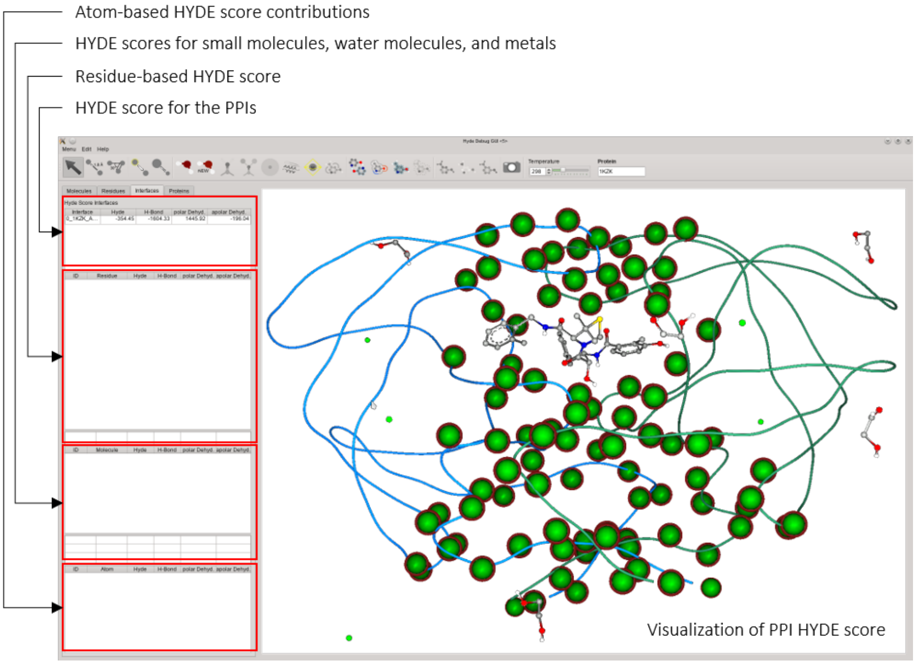
ID	Molecule	Hyde	H-Bond	polar Dehyd.	polar Dehy
802	EDO_B_802	-6.84	-27.26	23.17	-2.75
801	EDO_A_801	-4.28	-25.74	23.17	-1.71
905	CL_A_905	-1.41	0.00	0.00	-1.41
902	CL_A_902	-1.05	0.00	0.00	-1.05
701	JE2_A_701	-25.60	-80.15	81.12	-26.56
803	EDO_A_803	-5.10	-25.94	22.05	-1.21
804	EDO_A_804	-5.28	-25.79	23.01	-2.50
805	EDO_A_805	-6.27	-26.79	22.90	-2.38
901	CL_B_901	0.00	0.00	0.00	0.00
Total	Molecules	-57.90	-211.67	195.42	-41.65
Total	Waters	-312.00	-678.95	366.95	0.00
Total	Metals	0.00	0.00	0.00	0.00

ID	Atom	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.
5	CB	-0.36	0.00	0.00	-0.36
6	CG	-1.18	0.00	0.00	-1.18
7	CD	-0.70	0.00	0.00	-0.70
3	C	-0.03	0.00	0.00	-0.03
4	O	-2.04	-13.63	11.59	0.00
2	CA	0.00	0.00	0.00	0.00
1	N	0.00	0.00	0.00	0.00

Figure C.9: HYDE Debug GUI protein tab.

All protein-protein interfaces formed by two or more protein chains are scored with HYDE$_{protein}$ (Figure C.10). Their predicted affinities as well as the single amino acid contributions can be analyzed in the *Protein-Protein View*.

The last view is for organizational reasons only (Figure C.11). This allows different protein complexes to be handled simultaneously without the need to re-load them. All protein complexes once loaded into the HYDE Debug GUI are stored and can be selected for further analysis.


Atom-based HYDE score contributions
HYDE scores for small molecules, water molecules, and metals
Residue-based HYDE score
HYDE score for the PPIs
Menu
Edit
Help
Temperature
298
Protein
1K2K
Molecules
Residues
Interfaces
Proteins
Visualization of PPI HYDE score

Hyde Score Interfaces

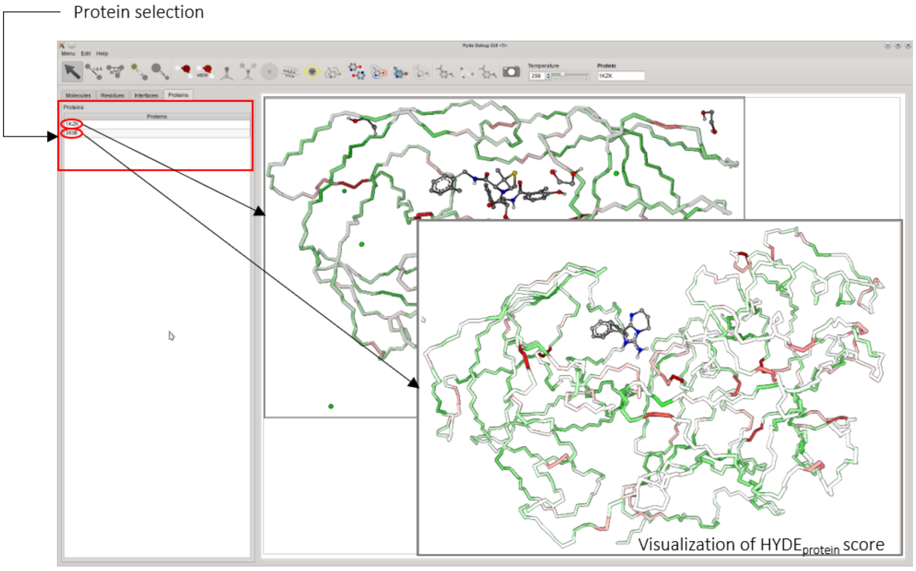
Interface	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.
0_1K2K_A...	-354.45	-1604.33	1445.92	-196.04

ID	Residue	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.

ID	Molecule	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.

ID	Atom	Hyde	H-Bond	polar Dehyd.	apolar Dehyd.

Figure C.10: HYDE Debug GUI protein protein tab.

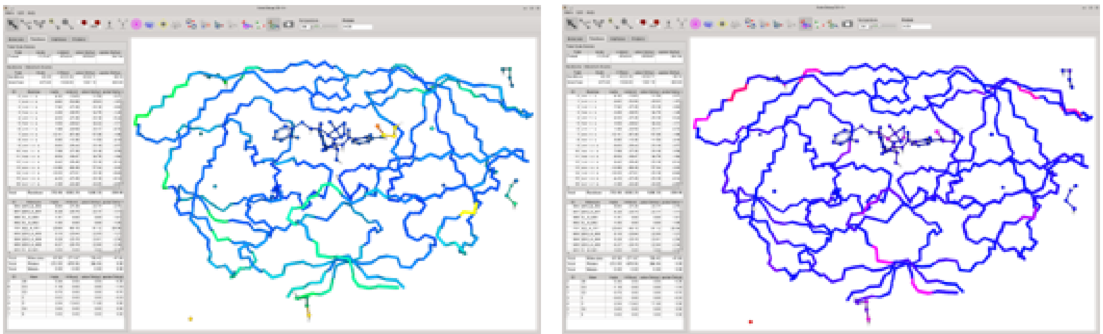


Proteins

Proteins
1K2K
3IOB

Figure C.11: HYDE Debug GUI protein selection tab.

In addition to the HYDE score visualization, the HYDE Debug GUI enables the user to analyze geometric arrangement by measuring distances and angles, as well as structural quality criteria (Figure C.12). The B factor coloring displays the B factors of each atom as annotated in the PDB file (Figure C.12a). Electron density maps can be displayed as well as the derived EDIA$_m$ values for each atom (Figure C.12b).



(a) B factor coloring for the protein backbone.

(b) EDIA$_m$ coloring for the protein backbone.

Figure C.12: Visualization of quality criteria in the HYDE Debug GUI.

The identification of free space can be visualized and water molecules can be placed using the water placement procedure described in this dissertation. Further visualization options comprise the molecular surface of the active site, the surface points of each amino acid or ligand separately, and the coordination of metals.

Parameter Selection

Diverse parameters used for the HYDE calculation can be adjusted by the user (Figure C.13). All parameters are briefly described below:

- **Settings mode**
  - Use SeeSAR settings: Applies the parameters and protein preparation steps used in SeeSAR[d]
  - Manually assign settings: Enables all of the following parameter settings.
- **HYDE settings**
  - Protoss: Optimization of the hydrogen bond network including protonation and tautomers.
  - Keep ligand protonation/tautomers: Keeps the input protonation/tautomers of the ligand.

[d] SeeSAR is a software tool for structure based design developed by BioSolveIT: https://www.biosolveit.de/SeeSAR/.

  - Keep protein protonation/tautomers: Keeps the input protonation/tautomers of the protein.
  - Active site radius: Selection of the radius of the active site for protein-ligand, protein, and protein-protein calculations.
  - Include PDB waters: Selection of water molecules from the input file or predicted by the water placement procedure.
    * None: All water molecules are deleted.
    * Relevant: Water molecules with either three potential hydrogen bond contacts to protein atoms and metals or one potential hydrogen bond to protein and another one to ligand atoms or three potential hydrogen bonds to ligand atoms are selected.
    * All: All water molecules are considered during HYDE calculations.
    * 3-protein bound: Only water molecules with potentially three contacts to protein atoms are kept for HYDE scoring.
    * 3-contacts: Only water molecules with potentially three contacts to protein or ligand atoms are kept for HYDE scoring.
    * Place waters: Water molecules are placed according to the placement procedure. All placed water molecules are integrated into HYDE scoring.
  - Dehydration penalty weight: Weight for the polar dehydration term of the HYDE scoring function $\Delta G_{dehydration}$
  - Hydrogen bond weight: Weight for the saturation term of the HYDE scoring function $\Delta G_{saturation}$
  - Internal H-bond threshold: Geometric quality threshold (f_{dev}) for the recognition of a hydrogen bond interaction.
  - Exposed term threshold: Penalty for exposure of apolar ligand atoms.
  - Use dehydration probability: Usage of the dehydration probability (p_{dehyd}) for HYDE calculation. If the dehydration probability should not be used the following discrete decision is made:

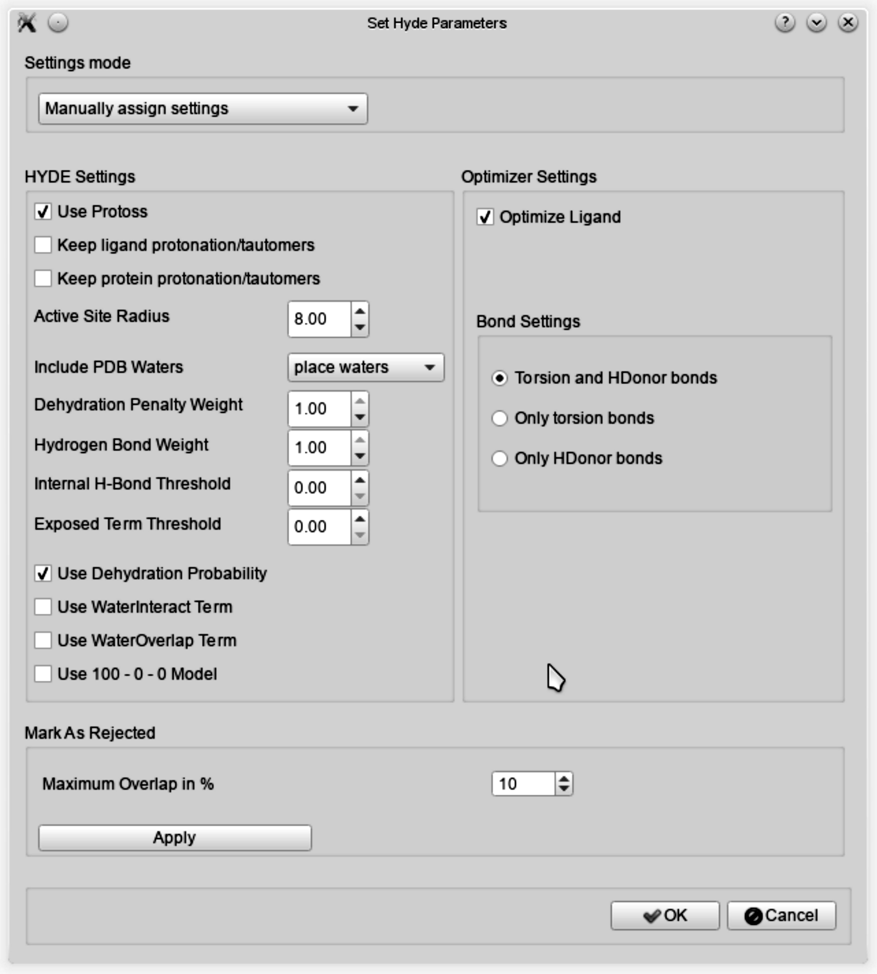
$$p_{dehyd} = \begin{cases} 0 & \text{if } p_{dehyd} < 0.5 \\ 1 & \text{if } p_{dehyd} \geq 0.5 \end{cases} \tag{C.1}$$

  - Use WaterInteract term: Inclusion of the WaterInteract term.
  - Use WaterOverlap term: Inclusion of the WaterOverlap term.
  - Use 100-0-0 model: Selection of 100-0-0 model or f_{dev}-based model.

- **Optimizer Settings**
  - Optimize Ligand: Activation of the geometric optimization of the ligand with GeoHYDE.
  - Bond Settings: Selection of bond parameters for optimization.

  - * Torsion and H-Donor bonds: Optimization of torsion angles as well as terminal hydrogen bond donors.
  - * Only torsion bonds: Only torsion bonds are rotatable during optimization.
  - * Only H-Donor bonds: Only terminal hydrogen bond donors are rotatable during optimization.

- **Mark as rejected**
  - Maximum Overlap in %: Visual indication of clashing molecules. The overlap is calculated using the van der Waals radii of the atoms. The percentage indicates the overlap of the van der Waals radii.



Set Hyde Parameters

Settings mode

Manually assign settings

HYDE Settings

☑ Use Protoss

□ Keep ligand protonation/tautomers

□ Keep protein protonation/tautomers

Active Site Radius 8.00

Include PDB Waters place waters

Dehydration Penalty Weight 1.00

Hydrogen Bond Weight 1.00

Internal H-Bond Threshold 0.00

Exposed Term Threshold 0.00

☑ Use Dehydration Probability

□ Use WaterInteract Term

□ Use WaterOverlap Term

□ Use 100 - 0 - 0 Model

Optimizer Settings

☑ Optimize Ligand

Bond Settings

● Torsion and HDonor bonds

○ Only torsion bonds

○ Only HDonor bonds

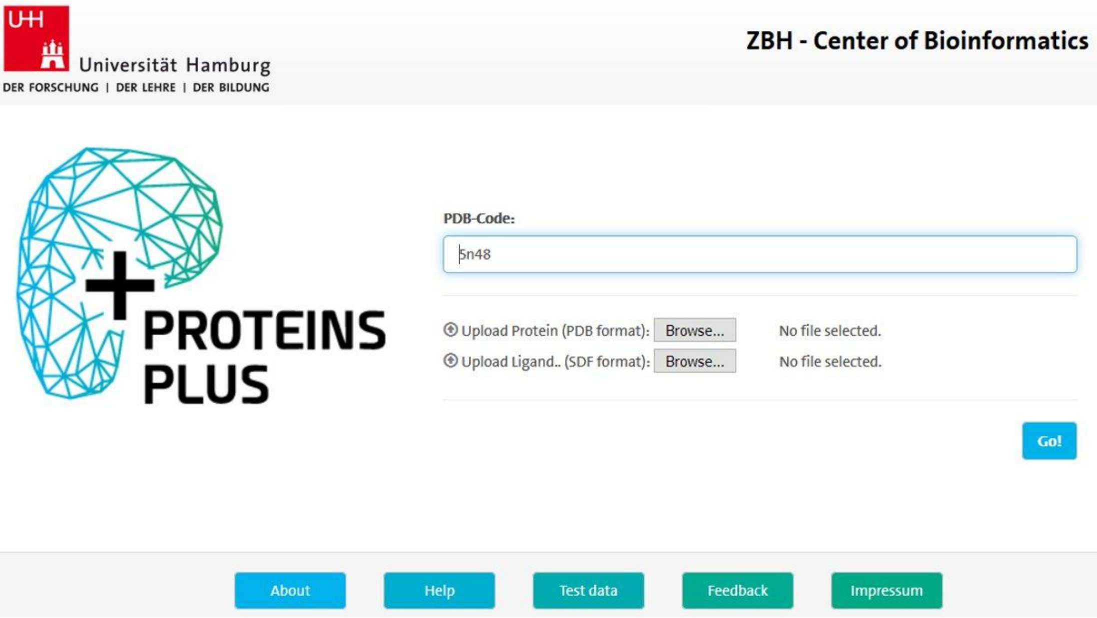
Mark As Rejected

Maximum Overlap in % 10

Figure C.13: HYDE Debug GUI parameter dialog.

C.2.3 HyPPI

HyPPI was developed for the automatic classification of protein-protein interfaces into crystal artifact or transient or permanent interface. Its integration in the Protein*Plus* web server allows the automatic retrieval of a protein structure from the PDB and its classification (Figure C.14).



PDB-Code: 5n48

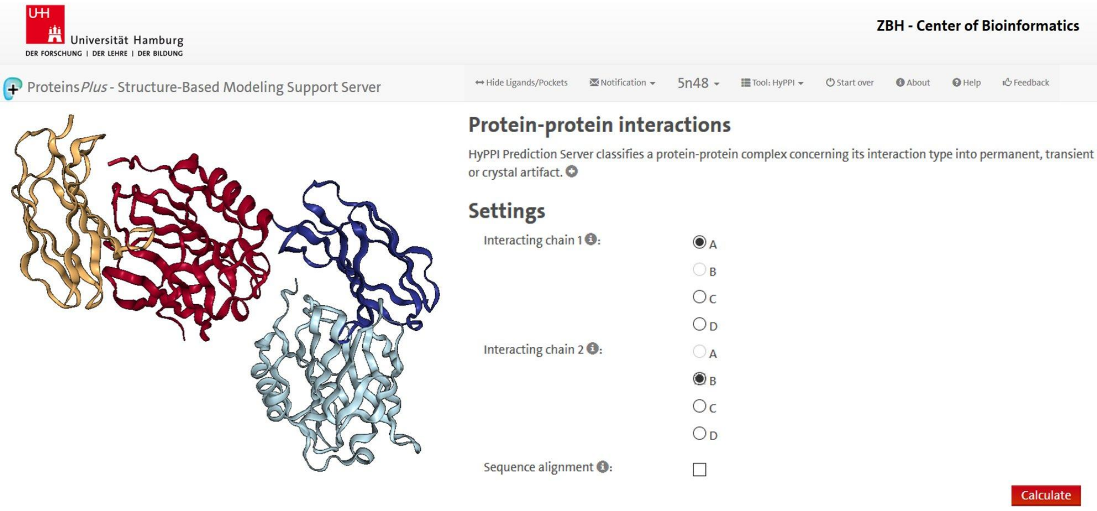
Upload Protein (PDB format): Browse... No file selected.

Upload Ligand.. (SDF format): Browse... No file selected.

Figure C.14: Proteins*Plus* interface; Example complex: Anticalin with a tumor antigen (PDBid 5n48).

The user can define the protein chains for which a classification should be performed (Figure C.15). Theoretically, multiple chains can be selected to form one interface. However, this is currently not implemented on the Protein*Plus* web server.

Once the classification is finished, which usually takes less than a minute, the probability for the interface to be a crystal artifact, a transient or permanent interaction are given (Figure C.16).


Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG
ZBH - Center of Bioinformatics
ProteinsPlus - Structure-Based Modeling Support Server
Hide Ligands/Pockets
Notification
5n48
Tool: HyPPI
Start over
About
Help
Feedback

Protein-protein interactions

HyPPI Prediction Server classifies a protein-protein complex concerning its interaction type into permanent, transient or crystal artifact.

Settings

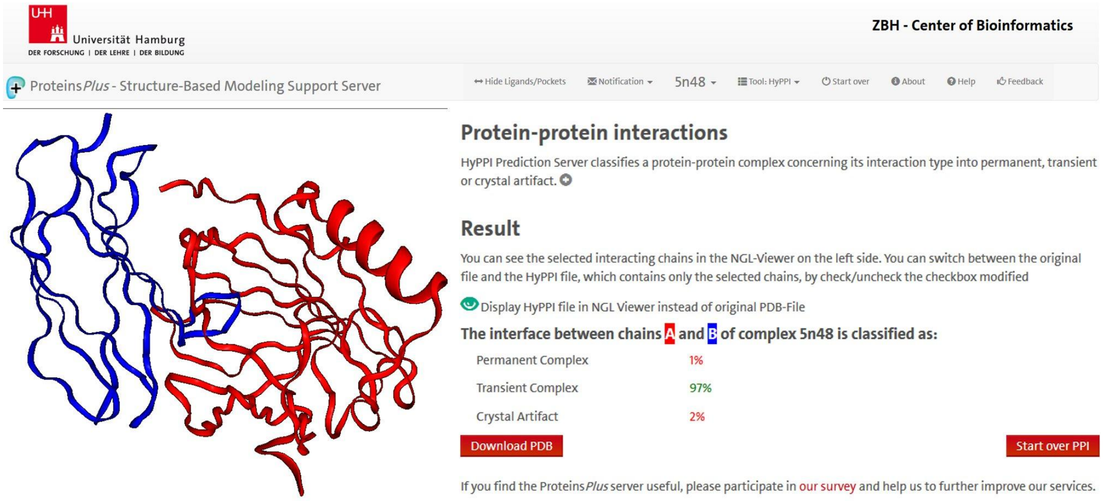
Interacting chain 1: ● A ○ B ○ C ○ D

Interacting chain 2: ○ A ● B ○ C ○ D

Sequence alignment: □

Calculate

Figure C.15: Selection of protein chains for their classification with HyPPI on the Proteins*Plus* web server; Example complex: Anticalin with a tumor antigen (PDBid 5n48).


Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG
ZBH - Center of Bioinformatics
ProteinsPlus - Structure-Based Modeling Support Server
Hide Ligands/Pockets
Notification
5n48
Tool: HyPPI
Start over
About
Help
Feedback

Protein-protein interactions

HyPPI Prediction Server classifies a protein-protein complex concerning its interaction type into permanent, transient or crystal artifact.

Result

You can see the selected interacting chains in the NGL-Viewer on the left side. You can switch between the original file and the HyPPI file, which contains only the selected chains, by check/uncheck the checkbox modified

Display HyPPI file in NGL Viewer instead of original PDB-File

The interface between chains A and B of complex 5n48 is classified as:

Permanent Complex	1%
Transient Complex	97%
Crystal Artifact	2%

Download PDB

Start over PPI

If you find the Proteins*Plus* server useful, please participate in our survey and help us to further improve our services.

Figure C.16: Results of the HyPPI classification on the Proteins*Plus* web server; Example complex: Anticalin with a tumor antigen (PDBid 5n48).

D
Journal Articles

D.1 Published Articles

Supporting Biocatalysis Research with Structural Bioinformatics.

[D1] Schneider, N.; Volkamer, A.; **Nittinger, E.**; Rarey, M. Supporting Biocatalysis Research with Structural Bioinformatics. In Applied Biocatalysis: From Fundamental Science to Industrial Applications; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2016; pp 71–100.

5
Supporting Biocatalysis Research with Structural Bioinformatics

Nadine Schneider, Andrea Volkamer*, Eva Nittinger, and Matthias Rarey*

5.1
Introduction

Computer methods have found their way to almost all fields of academic and industrial research by now. Especially, intricate design processes cannot be tackled without the use of specific software tools due to their inherent complexity. Although computational tools have been an integral component of most natural sciences disciplines over a long period of time, their application in life sciences to answer biological, biochemical, pharmaceutical, or biotechnological questions is still relatively sparse. This may be due to the complexity of the engaged systems. In many cases, the incomplete comprehension of the underlying biological and physicochemical processes constitutes an enormous challenge, which makes the generation of accurate theoretical models very difficult. In spite of this, many computational methods exist that try approaching these problems and already support experimentalists in life science research. This chapter focuses on the application of software tools to assist biocatalysis research. First, questions that can be addressed by computational methods, including an overview of methods currently deployed to biocatalytical problems, are identified. In the second section, novel computational methods are introduced, which have been developed for the analysis and comparison of protein binding pockets and the estimation of energetic contributions of protein-protein and protein-ligand interactions, respectively. In the third section, applications exemplifying the benefit of these novel methods for biotechnological research are given. Finally, a conclusion is drawn and future directions are discussed.

5.2
Computational Tools to Assist Biocatalysis Research

The discovery and efficient yield of biocatalysts in (bio-)technological processes is the central question in biocatalysis research. One of the challenges is to find

*A.V. and N.S. contributed equally to this work.

Applied Biocatalysis: From Fundamental Science to Industrial Applications,
First Edition. Edited by Lutz Hilterhaus, Andreas Liese, Ulrich Kettling, and Garabed Antranikian.

or optimize biocatalysts for specific processes. This objective goes along with the need to construct enzymes with new or enhanced catalytic properties. Thereby, the optimization of characteristics of the enzyme like temperature-, pressure-, or pH-stability often play key roles. Usually, experimental methods such as directed evolution are employed for this. Those are time- and cost-intensive approaches, based on several mutagenesis cycles combined with efficient screening or selection [1]. A variety of computational tools have been developed in the last years to assist directed evolution and de novo design of new enzymes; a comprehensive overview can be found in the work of Damborsky and Brezovsky [2, 3].

Besides the engineering of proteins, the identification of new enzyme classes from various species catalyzing specific reactions is another objective in biocatalysis research. Due to structural genomics projects and advances in crystallization techniques, nowadays, protein structures are elucidated before anything is known about their function. Experimentally determining the function of enzymes is a complex process, usually starting with the screening of known substrate collections; similar to finding a needle in a haystack. For this purpose, also several *in silico* methods have been developed to support the prediction of the protein's function as well as its properties considering the binding of substrates. In the following, computational methods related to the fields of *de novo* design, bioinformatics, and molecular modeling are presented in order to give an impression of the wide range of applications of those in questions occurring in biocatalysis research.

5.2.1
Computational Tools for Protein Engineering

In the process of protein engineering, the following questions may arise:

- Which residues should be modulated to optimize the enzyme activity or selectivity? And which are the best substituents for these residues?
- Up to which temperature is the enzyme stable and which mutations can lead to a better thermostability?
- Which residues are most important for a protein-protein interaction?

These and other questions considering the design of proteins can be addressed by a variety of computational tools, which are summarized in this section. Elaborate software suites for the *de novo* design of binding sites such as ROSETTA [4] or ORBIT [5] allow to construct virtually new enzymes that catalyze nonbiological reactions [6, 7]. A recently published tool PocketOptimizer [8] can be either used to optimize the active site of a protein concerning the binding affinity of small compounds or to establish the binding of new compounds by virtually mutating binding site residues. Besides these *de novo* design methods, quantitative structure-function or sequence-activity relationship analysis can be used as a computational tool to predict promising mutations concerning specific functional modulations [9, 10]. Here, a statistical model is derived from mutated enzymes using a set of structural and physicochemical properties of the amino acids and the activity of these enzymes. These analyses enable the classification and the prediction of beneficial, neutral, or disadvantageous mutations.

Although most mutations leading to functional enhancement of the enzyme were located in the binding site, other promising approaches exist that attempt to optimize ligand exchange pathways to improve the kinetics. The computational tool CAVER [11, 12] helps to find pathways within the protein that connect a buried active site with the solvent. If the three-dimensional (3D) structure of a target protein is resolved, CAVER can be used to identify important residues in these pathways. An application of this tool to redesign dehalogenase access pathways for degrading toxic substances has resulted in a 32 times higher activity of this dehalogenase toward a nonnatural substrate [13].

Furthermore, when analyzing protein-protein interactions, modulations of protein surface residues are of special interest. The COMBINE model [14] estimates residue-wise binding energy contributions of both interacting partners and, thereby, allows the identification of critical residues for binding. Another concept, called computational alanine scanning, estimates the specific contribution of a residue by virtually substituting it by an alanine [15, 16]. This approach is well suited to study protein-protein interactions as well as protein stability and can be combined with different energy functions.

Another important aspect in protein engineering is the thermostability of proteins. This can be analyzed using bioinformatics methods such as sequence alignment of homologous proteins to find, for example, the so-called ancestral mutations. Thereby, it is assumed that enzymes originate from a more thermostable but promiscuous ancestor and have been specialized later through evolution [17, 18]. A summary of all methods presented in this section is listed in Table 5.1.

5.2.2
Computational Tools for Function Prediction and Analysis of Enzymes

During the characterization of an enzyme with unknown function, some of the following questions may emerge:

- Which enzymatic class does the protein belong to?
- Do low-molecular-weight compounds (200–700 Da) bind to the protein?
- Which class of low-molecular-weight compounds will preferably bind to the protein? What are the physicochemical properties of such compounds?
- Can the function of a protein be inhibited or activated by a low-molecular-weight compound?
- Are there structurally or functionally critical waters involved in the binding process?
- Which other compounds in the bioassay may possibly interfere with the binding of the substrate to the enzyme?
- What is the bioactive binding mode of the substrate?
- Is a potential protein-protein contact deposited in the crystal structure the biologically relevant assembly?

Table 5.1 Summary of computational methods and tools for protein engineering.

Technique	Applications/goals	Computational tools/methods
De novo design of binding pockets	Design new enzymes that catalyze non-biological reactions	ROSETTA [4]
	Optimize binding pocket with respect to substrate binding affinity	ORBIT [5]
	Finding of non-natural substrates	PocketOptimizer [8]
Quantitative structure-function and sequence-activity relationships	Prediction of mutations concerning functional modulations	ProSAR [9, 10]
Modulation of binding pocket entrance	Optimization of ligand exchange pathways, identification of important residues	CAVER [11, 12]
Analysis of protein-protein interactions	Identification of critical residues for binding	COMBINE [14]
Computational alanine scanning	Analysis of energetic contributions of single residues	Alanine scanning [15, 16]
Sequence alignment of homologous proteins	Optimization of, for example thermostability	Analysis of ancestral mutations [17, 18]

Several computational tools already exist that focus on analyzing, for example the binding mode of a compound, the properties of the active site, or possible functions/substrates of an enzyme. In most cases, those tools depend on the availability of the 3D structure of the protein of interest. Due to advances in crystallization techniques, a growing number of protein structures are nowadays elucidated yielding large structural data pools. The freely available RCSB Protein Data Bank (PDB) [19], for example, contained more than 117000 structures in early 2016 and has shown an exponential growth since its launch. Using this abundant source of information, valuable insights about structure-function relationships of proteins can be derived with the help of computational tools.

Molecular docking methods (see for example [20] for a current review) are able to predict the bioactive binding mode of compounds in a protein binding site and to distinguish between compounds that will or will not bind to a protein. Using these methods with a set of potential substrates, insights into substrate specificity of enzymes have been successfully gained [21–25]. If the function of an enzyme is unknown, although its structure has been elucidated, docking approaches have already been used to propose potential substrates that reveal hints about its concrete function [26]. Protein function prediction is also possible by analyzing different properties of the complete protein or its binding pocket and relating these to the features of already characterized enzymes [27–32]. Furthermore, the analysis of molecular interactions between an enzyme and the substrate is essential to gain a deeper insight into the mode of action. Computational scoring functions have been designed for this purpose and are successfully applied in pharmaceutical research (reviewed in [33]).

To characterize the function of an enzyme, it is also important to know the natural biological assembly. In crystal structures, protein-protein interactions may be artificially formed due to the regular crystal lattice and often the biologically relevant multimeric state of the active enzyme is not obvious [34]. Several computational methods have evolved to distinguish between biological assemblies of proteins and crystal artifacts (see, for example, [35, 36]).

The mentioned approaches and success stories constitute only a subset of the available structure-based computational approaches, which can be, and partially already have been, successfully applied to biotechnological research. Nevertheless, few of these tools have been optimized toward solving biotechnological questions and many challenges remain. For a more detailed overview of the field of structural bioinformatics, the interested reader is referred to the following books [37–39].

In the next section, novel computational approaches are presented that address biocatalysis questions ranging from the analysis of enzyme binding sites and functional prediction of proteins to considerations of protein stability.

5.3 From Active Site Analysis to Protein Stability Considerations

Novel *in silico* approaches, presented in this section, were developed in close cooperation with industrial partners, thereby allowing to concentrate on urgent problems in biocatalysis research. These novel methods comprise tools for the structural and functional analysis of enzymes as well as approaches to assist protein engineering tasks such as optimizing protein stability.

One of the major objectives for the development of the novel approaches was enabling a comprehensive understanding of enzyme functions and properties, especially if nothing except the sequence and the structure of the protein is known beforehand. Focusing on structure, such analysis includes detecting potential binding pockets on the protein surface, characterizing them by structural and physicochemical descriptors and, finally, incorporating these descriptors for functional protein classification. For example, predicting the potential of an enzyme to be modulated or inhibited by low-molecular-weight compounds becomes possible based on these binding site descriptors.

Furthermore, the adaptation and improvement of the molecular docking tool LeadIT [40] enables predicting the natural substrate of an enzyme or finding new substrates with higher activity and, thus, better yield. In addition, docking substrates from known enzymes of specific classes into an enzyme binding site of unknown function can be used for functional classification.

In the context of enzyme optimization, that is the introduction of non-natural mutations, consideration of energetic and stability aspects is essential. Hence, a reliable estimation of interaction energies of protein-ligand and protein-protein complexes is required to allow systematic optimization and mutation of enzymes. For this purpose, the HYDE scoring function [41, 42], which has originally been developed to assess the interactions between ligands and proteins, has been adapted to this new application scenario.

In the following, the rational for developing the respective approaches together with the underlying methodology is discussed. First, the method for the detection and analysis of active sites is described. Second, the usage of derived pocket descriptors for classification of proteins, for example function annotation, is introduced. Furthermore, a short description of how to incorporate docking in an automated manner for substrate-specific functional annotation is given. Finally, the incorporation of the scoring function HYDE to assess the energetic contribution of molecular interactions in protein-ligand as well as protein-protein complexes is presented.

5.3.1
Computer-Aided Active Site Analysis of Protein Structures

The 3D structure of a protein is the key to its function. The formation of a protein-ligand complex, and thus, the completion of the biological function of a protein largely depends on the complementarity of the two binding partners. Protein and ligand have to adapt and fit to each other, similar to a key in a lock [43], demanding a structural as well as physicochemical match of the properties of the two binding partners.

As mentioned earlier, more than a hundred thousand 3D protein structures are currently publicly available. Although 3D structures are already used in biotechnological processes for the analysis of single enzymes, less effort has so far been undertaken to detect patterns from the whole data pool. Thus, automatic tools are needed to extract and categorize information from this data flood and to transfer this information to so far uncharacterized enzymes. Such knowledge-based transfer has long been used in sequence-based analysis, due to the fact that the sequence is usually known before its structure. Clearly, comparing the sequences of proteins will reveal information about the potential function of a protein based on high sequence similarity. Nevertheless, when this high similarity is missing, structure-based methods can still reveal similarities between distantly related proteins [44]. Furthermore, structural comparisons give insights into the spatial arrangement of key residues of an enzyme and can help to compare proteins on a functional level, and even give hints about how to optimize the yield of a biocatalyst. The automatic detection of potential binding pockets in protein structures is not a new task. Many algorithms have been developed over the last two decades that can be applied for *ab initio* pocket prediction [45]. Some challenges nevertheless remain, mostly due to the fact that the universe of pocket shapes is manifold together with the natural motion of proteins. Thus, a protein binding pocket can be small or large, buried or open, deep or shallow, continuous or disrupted making the correct detection of a pocket and its boundary in an automatic manner challenging. A new pocket detection algorithm has been developed, called DoGSite, which addresses especially the question of deriving distinct pockets and subpockets and of finding a correct boundary definition to make future descriptor-based analyses as feasible as possible [46]. The DoGSite algorithm has been evaluated on several retrospective benchmark studies and convinced by its good performance in recovering the true ligand-binding site. Special attention has been turned on the boundaries, precisely, the volume of the predicted pockets and subpockets. Measures such as the overlap of the pocket and the ligand have been used to show that the predicted pockets define restrictive volumes that include the major part of the ligand while leaving as little empty space as possible. This especially holds for the calculated subpockets which are, therefore, well suited for descriptor derivation and pocket comparison. Starting from this representation, various shape and physicochemical properties can be calculated, for example, volume, depth, and hydrophobicity of a pocket. Using such well-defined descriptors for the binding sites and, thus, the centers of action of a protein, allows correlating protein structures with their functional class. Finally, with respect to process optimization, predictions about the potential ability of an enzyme to be modulated or inhibited by low-molecular-weight compounds can be predicted with this method.

5.3.1.1 DoGSite: Binding Site Detection and Derivation of Representative Binding Site Descriptors

Similar to other geometric approaches for pocket detection, DoGSite uses a grid representation of the protein, but in contrast to other algorithms it incorporates a difference of Gaussian (DoG) filter from the field of image processing for cavity detection and enables the prediction of more reasonable pocket boundaries. Furthermore, the new algorithm is able to separate detected pockets into subpockets allowing a more detailed analysis.

The procedure is straight forward: The 3D structure of a protein complex, including all protein atoms, their types and their locations, that is x-, y-, and z-coordinates in Cartesian space, is used as input. A Cartesian grid is spanned around the protein (Figure 5.1a). Subsequently, each grid point is scanned and assigned as occupied if it lies inside the van der Waals radius of any protein atom, otherwise as free (see Figure 5.1a, occupied = 1, free = 0). Next, the DoG filter is applied to the grid detecting invaginations on the protein surface where positioning of sphere-like objects is favored (Figure 5.1b). Thus, grid points with favorable DoG values are merged to subpocket cores (Figure 5.1c). In the final step, these subpockets are dilated toward the protein surface and, eventually, merged into pockets (Figure 5.1d).

Analyzing binding sites based on their shape and physicochemical features can generate valuable information for further design processes. For example, similarities between distantly related proteins can be explored by comparing their binding site features.

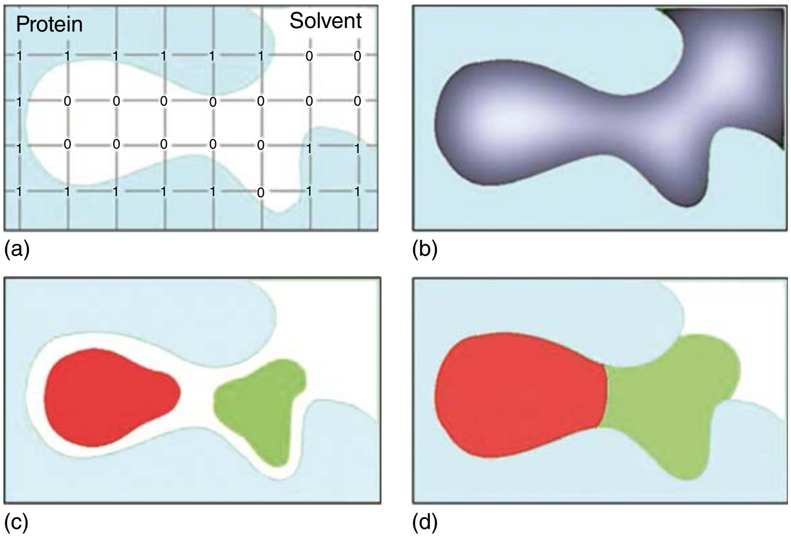


Figure 5.1 Schematic depiction of DoGSite's (sub)pocket detection. (a) Grid representation of the protein binding site (light blue). (b) Filtering of the grid using a DoG filter. (c) Merging of favorable grid points to subpocket cores. (d) Dilation to subpockets and merging to one pocket.

To describe the size of the cavity, the volume and the surface of the pocket are calculated. The depth of the pocket is described by the largest distance between any solvent-exposed pocket grid point and the most distant buried grid point. Furthermore, the exposure of the pocket can be described by the ratio of number of solvent-exposed grid points to the number of pocket-lining grid points. Finally, the shape of the pocket is mimicked by ellipsoids fitted into the pocket volume, thus, simplifying the shape as being something between a rod, a disk, and a sphere (see Figure 5.2).

The physicochemical properties of the pocket-lining atoms are equally important for ligand binding and are, therefore, also added to the set of pocket descriptors. The amino acid composition of the pocket-lining residues is calculated based on the respective type and then grouped by their physicochemical properties. Furthermore, functional groups, for example, hydrogen bond donor and acceptor atoms, as well as hydrophobic groups are detected and a hydrophobicity profile is calculated. In total, over 40 properties are collected, which can be used for the analysis and comparison of enzymes. A more detailed description of the method and the derived properties of the binding pockets can be found in [46].

5.3.1.2 DoGSiteScorer: Descriptor-Based Protein Classification

Structural descriptors, derived from the active site, allow to correlate protein structures with their specific function, family affiliation, or binding behavior. Thus, to assess the structure-function relationship of proteins, the above-mentioned descriptors are incorporated into two different classification methods.

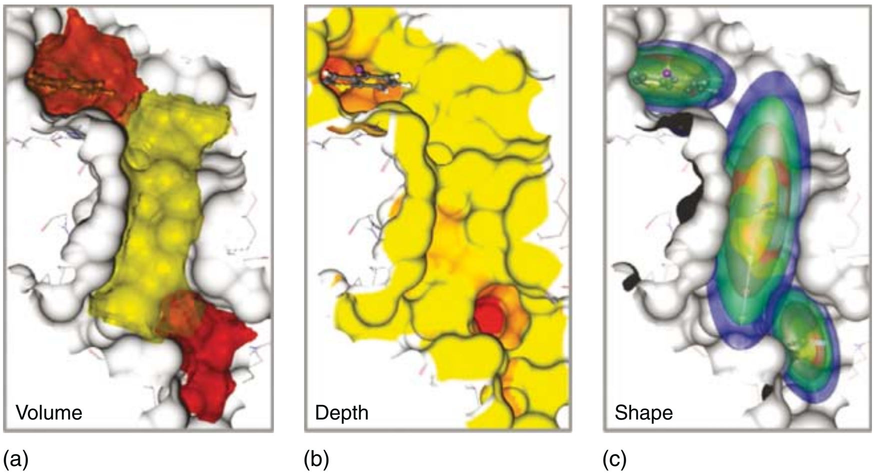


Figure 5.2 3D structures of the active site of urokinase-type plasminogen activator (PDB code 1c5q) including three exemplarily calculated descriptors. The protein surface is shown in gray. The co-crystallized ligand is depicted in ball-and-stick mode. (a) Volumes of the three subpockets sketched in orange, yellow, and red. (b) Depth of the pocket, color coding from yellow (solvent exposed) to red (buried). (c) Ellipsoidal shapes calculated for all three subpockets.

The first classification approach, a hierarchical clustering method, has been incorporated to group proteins by the similarity of their descriptor profiles. The analogy between two proteins is calculated based on the summed similarities or distances between the single descriptors of the respective proteins and normalized to a value between zero and one. Based on these values a cluster tree can be calculated holding information about the relationship between proteins or, more precisely, their binding sites (similar to a phylogenetic tree showing the evolutionary relationship among various entities). This approach has been applied in a mutation study and is further explained in the application part (see Section 5.4.3).

Nevertheless, in most cases, there is no simple linear relationship between one or several descriptors and the functional class of a protein. In such cases, other sophisticated machine learning techniques can be applied, which are more robust in assigning the correct class for nonlinear data. Therefore, in DoGSiteScorer an existing freely available implementation of a support-vector machine (SVM) [47] was chosen. Besides its robustness, this SVM can be applied to separate multiple groups, for example necessary to classify proteins into the six EC main classes. Furthermore, the classification is supported by a probability value. Based on the above-mentioned descriptors, active sites can be correlated to arbitrary classification scenarios, such as the prediction of the potential of an enzyme to be inhibited or activated by low-molecular-weight compounds or the function of an uncharacterized enzyme (more details about the application are given in Section 5.4.1).

The premise for these machine learning methods is a reliable and large data set, in which each protein is represented by a well resolved structure and a distinct assignment to a class. Having collected such a data set, the binding pockets for the respective protein structures are detected using the DoGSite algorithm and the set of descriptors is calculated. Usually, the data set is separated into training and test data. The training set is used to train the SVM method to optimally separate the data points from the different classes based on the precalculated descriptors. The independent test set is then employed to evaluate the prediction performance of the built SVM model. Once a model has been established, it can be queried with any new protein structure by computing the respective descriptors and afterward assigning, for example its ability to bind molecules of a specific type or its function in general. A more detailed description of the classification method is given in [48, 49].

5.3.2 Molecular Docking to Assist Functional Characterization of New Enzymes

As an alternative to the concept of pocket similarity, which is applied in the DoGSiteScorer to derive the function of a protein, molecular docking can help to identify the substrate of an enzyme and, thereby, its function. The LeadIT software suite [40], which includes the FlexX docking algorithm [50], has originally been evolved for computer-aided drug design. The idea of molecular docking is to place a small molecule of interest in the active site of a protein in order to identify the bioactive binding mode. Hence, docking algorithms are confronted with the lock-and-key problem, that is, how to best fit the small molecule (key) into the protein binding site (lock). Thus, the initial requirement is a strategy to place the ligand in the binding site, followed by the evaluation of the calculated binding mode using a scoring function. To assess the quality of a ligand binding mode, the scoring function estimates the respective free binding energy. Most scoring functions have been calibrated using a set of protein-ligand complexes with resolved crystal structures and experimentally measured affinities.

Since in nature proteins and their binding partners are flexible, the two binding partners are able to adapt their conformation to each other (known as induced-fit phenomenon [51]) to achieve a better fit. Modeling this flexibility is rather challenging. Thus, the first docking approach kept protein and ligand rigid [52], while subsequent approaches tried to capture ligand-flexibility (see, e.g., [50, 53, 54]), eventually, most recent tools also try to investigate protein flexibility (see, e.g., [55, 56]). In the FlexX docking algorithm (integrated in the LeadIT software), the ligand is treated in a flexible manner.

In order to apply molecular docking in biocatalysis questions, the LeadIT software had to be adapted. As mentioned earlier, scoring functions are calibrated on protein-ligand complexes, mostly on protein-inhibitor complexes, leading to a high dependency of the performance on the underlying complexes. In contrast to inhibitors, which exhibit a strong binding to the protein, substrates bind rather weakly to an enzyme. This is problematic since most scoring functions were only trained on strong binders and, therefore, often cannot correctly assess the binding mode of substrates. For this purpose, the HYDE scoring function (see next Section 5.3.3), which is not calibrated on experimental binding affinities, was incorporated into the LeadIT software to enable a reliable estimation of the binding energy of substrates. Another issue attributed to substrates is the importance of the correct protonation state. Hence, the docking tool must be able to handle different protonation states and tautomeric forms. For this reason, the tool Protoss [57, 58], which is able to find the best hydrogen bonding interaction network, was also included in the LeadIT software.

These major adaptations enable the LeadIT docking software to be used for functional classification and specificity predictions. An open issue is the transition state substrates adopt during the catalytical process. This high-energy state of a molecule is usually not modeled in docking tools, but can be incorporated by generating these states beforehand [26]. In the application part of this chapter, two successful examples using the LeadIT docking software in function as well as specificity prediction will be shown without especially considering transition states of substrate (see Sections 5.4.4 and 5.4.6).

5.3.3
Energetic Estimation of Protein-Ligand and Protein-Protein Interactions

The reliable estimation of binding free energy between two biomolecules is a prerequisite for the understanding and modeling of biological processes. Almost all life sciences – whether dealing with biotechnological process optimization or the development of new pharmaceuticals – will benefit from the solution of this problem. A wide variety of issues could be tackled, ranging from optimization of the selectivity of an enzyme over systematic determination of mutations to enhance the thermostability of a protein to the correct assignment of the biological function of an enzyme by identifying the natural substrate. Several computational approaches exist trying to reliably estimate the binding affinity of biological complexes. Elaborated methods to calculate the free energy such as quantum chemistry calculations or molecular dynamics simulations are time and resource consuming, allowing their application only on small systems or particular questions. Alternatively, scoring functions have been successfully applied, mainly in pharmaceutical industry, to assess the binding affinity of a compound to a protein target [33]. These functions are mathematical expressions to estimate the energetic contribution of noncovalent protein-ligand interactions. Herein, it is assumed that independent interaction contributions could be additively combined to calculate the total binding free energy. Normally, scoring functions rely on calibration strategies, which include experimentally measured binding affinities of protein-ligand complexes and their 3D structures. This induces a high dependency of the performance of a scoring function on the quality of the underlying data as well as on the types of complexes used for its calibration. Another challenge is the inclusion and description of destabilizing interactions which are rarely found in 3D structures. In general, the precise modeling of molecular interactions and balancing energetic contributions of different kinds of interactions is still a matter of research. Furthermore, the consideration of enthalpic and entropic contributions to binding free energy is challenging, given that the main part of entropy is attributed to the surrounding solvent molecules. Hence, to solve the problem of binding free energy estimation many different aspects have to be considered.

In the past decade, the scoring function HYDE has been developed for protein-ligand complexes, without using experimental affinity data for calibration [41, 42, 59]. The HYDE scoring function has so far been applied successfully to a variety of pharmaceutically relevant questions showing good results in line with other highly parameterized scoring functions in the field [42, 59]. In the following sections, the concept behind the HYDE scoring function is described and new potential applications of HYDE in the biotechnology area are presented.

5.3.3.1 The Concept behind the HYDE Scoring Function

The theoretical concept behind the HYDE scoring function describing the saturation of the hydrogen bond network in liquid water has been developed by Lange and Klein [60]. The fraction of satisfied hydrogen bond functions $F_{\mathrm{sat}}(T)$ at a certain temperature T could be derived from the thermodynamic cycle of water (see Figure 5.3) under the following assumptions: First, in hexagonal ice crystals, the water molecules are completely satisfied forming four hydrogen bonds with their neighbors. Second, in a vaporous state, all four hydrogen bonds are broken. Finally, the energy that is needed to break all four hydrogen bonds can be deduced as the sum of three enthalpic terms: the enthalpy of fusion, the enthalpy of heating up water from 273 K to 373 K, and the enthalpy of evaporation. This results in 54.18 kJ mol^{-1} (see Figure 5.3).

Based on these findings, the temperature-dependent fraction of unsatisfied hydrogen bond functions $F_{\mathrm{unsat}}(T)$ can be estimated by dividing the actual energy of the system at temperature T by the total energy needed to transfer the water

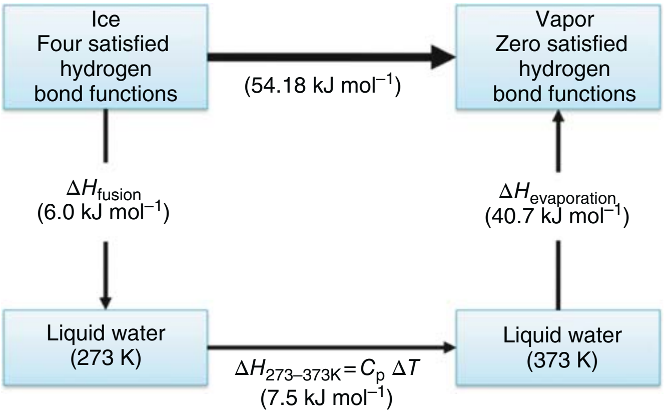

Ice
Four satisfied hydrogen bond functions
(54.18 kJ mol^{-1})
Vapor
Zero satisfied hydrogen bond functions
$\Delta H_{\mathrm{fusion}}$
(6.0 kJ mol^{-1})
$\Delta H_{\mathrm{evaporation}}$
(40.7 kJ mol^{-1})
Liquid water (273 K)
$\Delta H_{273\text{–}373\mathrm{K}} = C_p\,\Delta T$
(7.5 kJ mol^{-1})
Liquid water (373 K)

Figure 5.3 Thermodynamic cycle of water, C_p = heat capacity of water.

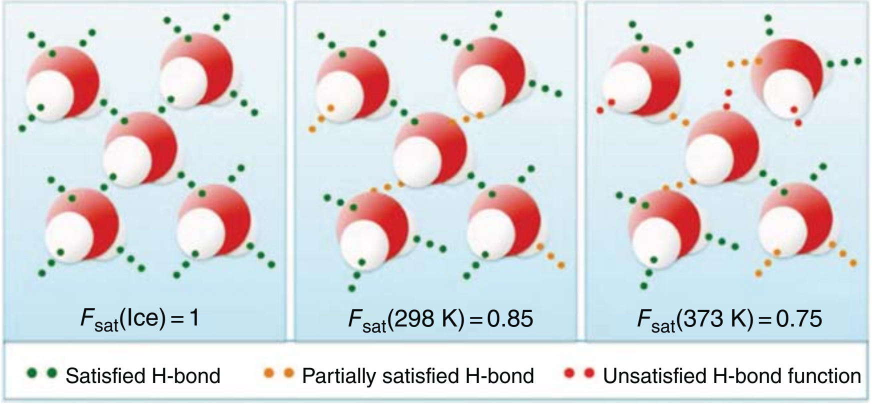


Figure 5.4 Saturation factor $F_{sat}(T)$ at different temperatures.

molecules into the vaporous state (see Equation 5.1). The remaining fraction represents the amount of satisfied hydrogen bond functions $F_{sat}(T)$ in bulk water at a given temperature:

$$F_{unsat}(T) = \frac{\Delta H_{fusion} + C_p \cdot (T - 273\,\text{K})}{\Delta H_{fusion} + C_p \cdot (373\,\text{K} - 273\,\text{K}) + \Delta H_{evaporation}}$$

$$F_{sat}(T) = 1 - F_{unsat}(T) \tag{5.1}$$

Figure 5.4 gives an overview of the range of values the saturation factor $F_{sat}(T)$ can adopt at different temperatures. Binding affinities were usually measured at ambient temperature (298 K); here, the saturation factor F_{sat}(298 K) equals 0.85, which means that 85% of the hydrogen bonds of the solvent molecules were satisfied (assuming the solvent is an aqueous solution). The saturation factor is an important parameter, which is included in the HYDE scoring function (discussed later). Furthermore, using this theoretical concept, dehydration terms and values for idealized hydrogen bond functions could be derived. A more detailed description of this concept can be found in [60].

5.3.3.2 HYDE – Estimation of Hydrogen Bonding and Dehydration Energy

The HYDE scoring function models the basic concepts of binding. In the unbound state, protein and ligand are solvated in aqueous solution. To enable the binding process water molecules located in the binding pocket of the protein are displaced while those surrounding and interacting with the ligand are stripped off. Primarily, this leads to an unfavorable enthalpic contribution since hydrogen bonds of protein and ligand to water molecules are broken. Establishing new hydrogen bonds between protein and ligand may counterbalance this energy loss. In addition, hydrophobic moieties of ligand or protein, which have been in contact with water molecules beforehand, lead to an unfavorable energy given that they introduce a discontinuity in the water hydrogen bonding network. Removing these water molecules from the hydrophobic surfaces and releasing them to bulk water produces a gain in energy, the so-called hydrophobic effect

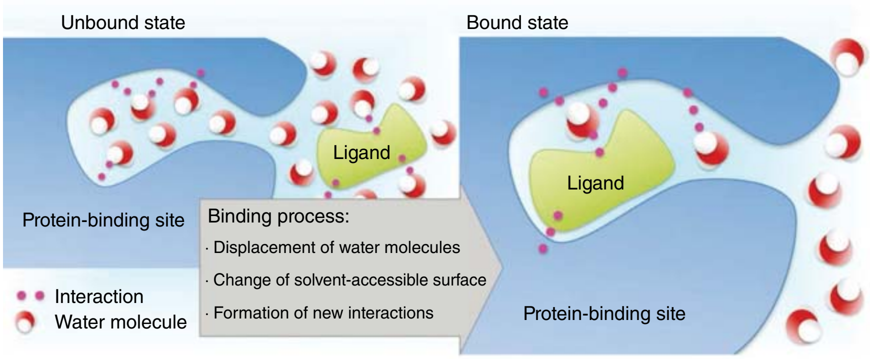


Figure 5.5 Schematic depiction of the binding process, modeled in the HYDE scoring function.

[60]. In HYDE, it is assumed that the main energetic contributions to the binding energy arise from the above-described processes. Hence, the interactions modeled in the HYDE scoring function are hydrogen bonding and the hydrophobic effect as well as the unfavorable contribution of dehydration of polar atoms (see Figure 5.5).

The equation of the HYDE scoring function consists of two terms: one to calculate the change in hydrogen bonding ($\Delta G_{\text{H-bonds}}$) and a second one to estimate the dehydration energy ($\Delta G_{\text{dehydration}}$) for every atom i in the protein-ligand interface (see Equation 5.2).

$$\Delta G_{\text{HYDE}} = \sum_{\text{atoms } i} \Delta G^{i}_{\text{H-bonds}} + \Delta G^{i}_{\text{dehydration}} \tag{5.2}$$

Both terms of the HYDE scoring function – the hydrogen bonding and the dehydration term – are derived from the Gibbs-Helmholtz equation ($\Delta G = -\text{RT}\ \ln(K)$). As equilibrium constant, an atom-based $\log P$ ($p\log P$) value is introduced in the hydrogen bonding as well as the dehydration term. The atom-based $p\log P$ increments were derived from experimental $\log P$ values (octanol-water partition data) of small molecules using multiple linear regression (MLR) analysis [61]. The energy contribution of a hydrogen bond in HYDE arises from the fact that statistically not all hydrogen bonds in bulk water are perfectly realized (discussed earlier). Thus, the energy for disrupting these weak hydrogen bonds is lower than that for ideal hydrogen bonds [60]. This phenomenon is integrated into HYDE by using the saturation factor $F_{\text{sat}}(T)$ (see Equation 5.1). A more detailed description of the HYDE scoring function is given in [42].

5.3.3.3 Estimation of Protein-Protein Interactions Using HYDE

The calibration strategy pursued in HYDE – the usage of octanol-water partition data of small molecules instead of binding affinities and crystal structures of protein-ligand complexes – prevents HYDE from being restricted to the estimation of binding energies in protein-ligand complexes. Furthermore, the general concept behind the HYDE scoring function allows assessing interactions between arbitrary molecules which take place in aqueous solution. In addition, the before-mentioned temperature-dependent saturation factor $F_{\text{sat}}(T)$ allows to include the temperature dependence of molecular interactions in HYDE, enabling the analysis of thermostability of proteins. These are the required foundations making HYDE applicable to biocatalysis questions such as the systematic optimization of enzymes, directed mutational analysis of proteins, or the improvement of selectivity.

To realize the scoring of protein-protein interactions within a protein structure and to assess whether the 3D structure is stable at different temperatures, the reference state used in the HYDE function has to be changed. Unlike the estimation of protein-ligand binding energy, in which the energy difference of unbound and bound state is calculated (see Figure 5.5), the energy gain between the unfolded and folded state has to be assessed to prove the stability of a protein structure. In the unfolded state, a residue is solvated and able to freely interact with the surrounding water molecules, whereas in the folded state, the residue is almost completely dehydrated and restricted to interactions with neighboring residues in the final 3D structure of the protein. Hence, for each residue of the protein, the difference in hydrogen bonding and dehydration energy between these two states is estimated using the HYDE function. Furthermore, the terms of the HYDE scoring function can be slightly modified to include the temperature-dependent saturation factor $F_{\text{sat}}(T)$ also in the dehydration term, identifying protein residues contributing more or less favorable to binding at elevated temperatures. To further assess the stability or the type of protein-protein interfaces found in multimeric protein complexes, the identification of all potential interfaces contained in the 3D structure of the protein complex is implemented. Afterward, the binding energy of all these interfaces is estimated using the HYDE scoring function.

These adaptations enable the investigation of the energetic effect of mutations, the analysis of stability of the protein structure at different temperatures, and the classification of protein-protein interfaces in biologically relevant assemblies and artificial crystal contacts.

5.4
Applying DoGSiteScorer and HYDE to Biocatalytical Questions

In the following, some exemplary applications of the computational tools described in Section 5.3 are presented. These examples comprise some of the biocatalysis questions, which were mentioned in Section 5.2 and which could be addressed using the newly developed computational approaches. In the first application, one of the most important questions in biocatalysis research, the annotation of the function of an enzyme, is addressed by the DoGSiteScorer SVM-based classification method. This question could also be tackled by molecular docking and is shown in a subsequent application study. Another challenge in biocatalysis research is the suggestion of potential mutation sites, which can be addressed incorporating the descriptor-based binding site comparison. Furthermore, the derivation of specific properties of the binding site and their usage to decide whether a binding pocket could be targeted by a low-molecular-weight compound are described. Subsequently, another important question, the prediction of competitive substrate inhibition by other compounds within the activity assay, is discussed. Finally, the last application study is the determination of the biologically relevant assembly of a protein-protein complex deposited in the PDB crystal structure using the HYDE scoring function. This is only an extraction of possible questions that could be answered by the developed computational tools. Further applications are discussed at the end of the chapter.

5.4.1
Enzymatic Function Prediction Using the DoGSiteScorer

Due to, for example, structural genomics projects and advances in crystallization techniques, nowadays protein structures might get elucidated before the function is fully characterized. Many sequence-based approaches for functional annotation are known. However, proteins without homologous sequences can still share functions and *vice versa* proteins with high sequence similarity can disagree in their functional duty. Several structural methods are available for function prediction like fold comparison (SCOP [62], CATH [63], eFold [64]), structural alignments (PAST [65], VAST [66]), descriptor-based comparison [27, 28] or structure-based (fragment) docking [26]. Despite this magnitude of methods, many protein structures are still deposited in the PDB with missing or wrong functional annotation. In a recent study, the number of nonredundant X-ray structures in the PDB with "unknown function" has been reported to be 2549 [67]. While more than half of the proteins could be reassessed by further investigations into UniProt Knowledgebase, sequence and fold similarity, a total of 1084 protein structures remained with "true" unknown function.

The DoGSiteScorer SVM-based classification approach has been advanced for answering biotechnological questions as the prediction of enzymatic function based on the EC classification scheme. To assist overcoming the remaining functional annotation lack, a new data set containing all proteins from the PDB with annotated EC number was generated [49]. Over 26 000 pockets, containing a bound ligand and fulfilling the implemented coverage quality criterion, were detected and used for the training and testing of the method. Based on the calculated global properties, these pockets were separated on different EC specification levels based on multiclass SVM models. The method especially convinced through the introduction of a stepwise classification into EC main class, subclass, and substrate-specific sub-subclass. With aid of this method, a deposited protein of unknown function or a newly elucidated protein can be classified step-by-step, thus predicting its potential EC class together with estimated confidence values for the respective annotation steps. As an example, a hypothetical protein TM0936 (PDB code: 1p1m) has been investigated in a retrospective study. The protein was clearly predicted as being a hydrolase (EC 3) and could further be classified into subclass EC 3.5. Finally, as substrate-specific classification a peptide amidohydrolase (EC 3.5.1.88) and an adenosine deaminase (EC 3.5.4.4) were found on the top ranks. This function prediction was in good agreement with other annotation methods and literature reports [26]. Although, the method may not give a unique answer to the functionality question, it can generate reasonable hypothesis of the protein function within seconds, which could, in a next step, be verified biochemically.

The DoGSiteScorer method for active site analysis has also been used for the classification of four selected enzyme families: lipases (LED), cytochromes (CYPs), thiamindiphosphate-dependent enzymes (ThDPs), and medium-chained dehydrogenases/reductases (MDRs) with a total of 943 structures [68]. The setup procedure of the classification models was the same as described in the previous paragraph, and the accuracy to annotate the correct enzyme family in a cross-validation study on this data set was 91%.

5.4.2 Docking-Based Functional Protein Classification

Besides the enzyme function prediction via descriptor analysis, the docking software LeadIT was employed for this task. The aim was to classify proteins based on docking scores of specific reference substrates. The idea for substrate-based function prediction is to count the quantity and quality of enzyme-class-specific substrates binding to an enzyme of interest. Thus, in this experiment, it was assessed which substrates from proteins of known function bind best to the proteins of unknown function.

The above-described enzyme data set containing the four enzyme families was adapted with focus on the bound substrates. All substrates were extracted; erroneous molecules and duplicates were removed, yielding a set of 189 substrates which were docked iteratively against the enzymes. This retrospective application was aimed to recover the function of an enzyme pretending that nothing about the function was known beforehand. The experiment is exemplified for a known lipase (considered as unknown protein, see Figure 5.6). In this test case, all extracted substrates have been docked against the target enzyme and the resulting docking poses have been sorted by docking score. Based on the experiment design, the substrates from known lipases should be accumulated on the top ranking scores, and indeed they did (as shown on the right-hand side of Figure 5.6). Thus, to predict the most likely function of the target enzyme, the docked poses have been ranked by their scores and a score histogram was calculated (see Figure 5.6, right). Clearly, the lipase substrates (actives) got better (lower) scores than the substrates from the other three classes (decoys). Repeating the experiment for all enzymes in the data set, a mean accuracy of 73.5% over all four classes was observed. In general, the function prediction via docking worked quite well; the performance was especially good for lipase recovering, whereas the performance on ThDP enzymes was rather poor. This implies potential problems in the preparation of the data

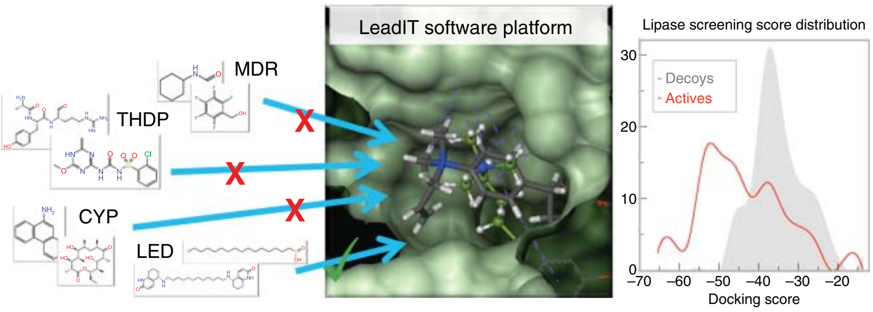


Figure 5.6 Exemplified function prediction pipeline via docking of substrates into the active site of a new enzyme. Left: Example substrates from four different enzyme classes. Middle: The binding site of an acetylcholinesterase (1acl, green surface), exemplarily with a co-crystallized ligand (DME) is depicted in light green and the docked ligand (EBW from 1e3q) in gray. Right: Histogram of the achieved docking scores. The red curve shows the scores of the actives (LED substrates) and the gray curve the scores of the inactives (MDR, ThDP, CYP substrates).

set, the choice of substrates, or the consideration of cofactors or ions. Furthermore, in some cases, the docking method may generate nonnatural binding modes leading to wrong assumptions. The results show that the LeadIt docking software is capable of generating information about the catalytical function and substrate specificity of an enzyme. A further example using docking to predict competitive substrate inhibition is presented in Section 5.4.5.

5.4.3 Predicting Potential Mutation Sites Using DoGSite and Molecular Modeling

The mutation of enzymes to optimize the substrate conversion is a major issue in biotechnology. Directed evolution and random mutations are only two examples of how to achieve enzymes with better yields. In addition, using computational tools for rational enzyme design can help to detect additional potential mutation sites. As one example, the conversion of an alditol oxidase via directed evolution into a glycerol oxidase was investigated [69]. Since directed evolution did not achieve the expected increase in glycerol activity, the assistance of DoGSite was consulted. A strategy was established to predict potential mutation sites based on binding site comparisons combined with molecular modeling.

The goal was to use typical binding properties of glycerol-binding enzymes as idea generator for mutations in the active site of alditol oxidases toward an enhanced glycerol activity. A PDB search for oxidases binding alditol and those oxidases in complex with glycerol yielded 210 glycerol oxidases and 5 alditol oxidases with known structure (PDB codes: 2vfr, 2vfu, 2vfv, 2vfs, 2vft). Next, to detect similarities and differences between their active sites, DoGSite was applied to find and describe the respective oxidase pockets. Subsequently, the proteins have been clustered based on the previously described similarity

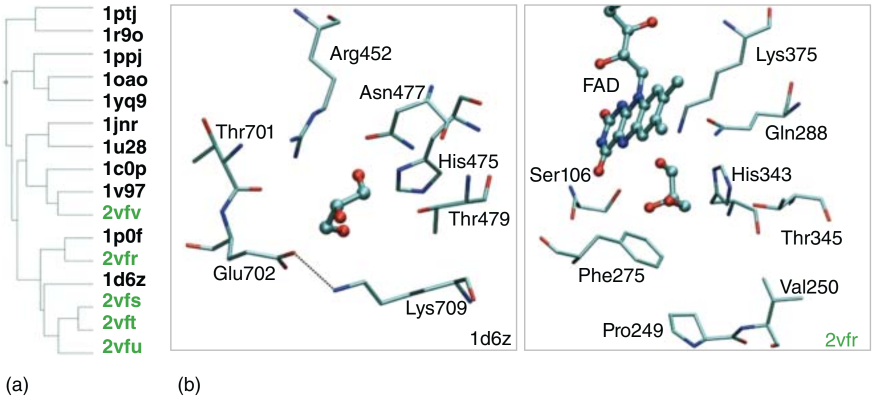


Figure 5.7 (a) Subset of clusters resulting from the descriptor-based binding site comparison of 210 glycerol oxidases and five alditol oxidases. The shown subtree includes the five alditol oxidases (green) and the most similar glycerol oxidases (black). (b) Comparison of the active sites of the glycerol-binding oxidase (1d6z) and the alditol oxidase (2vfr). Shown are amino acids that are in a similar position within both binding sites. A salt bridge between Glu702 and Lys709 is depicted as dotted line. Only C_1 to C_3 of xylitol in alditol oxidase (2vfr) are shown.

measure. The obtained clustering tree showed that the five alditol oxidases are very similar and ended up in one branch, together with 11 glycerol binders (see Figure 5.7a). Out of these 11 structures, the glycerol-binding oxidase (PDB code: 1d6z), which is most similar to four of the alditol oxidases, was chosen for further investigations. To exactly characterize the differences in these structures, the amine oxidase (1d6z) was superposed onto the alditol oxidase (2vfr) and analyzed using the molecular modeling software MOE [70]. Direct comparison allowed the detection a few residues, which could be mutated in the active site of the alditol oxidase to closer resemble the known active site of the glycerol-binding oxidase (see Figure 5.7b).

In 1d6z, the two amino acids Glu702 and Lys709 form a salt bridge. With respect to the carbon atoms, two equidistant amino acids, that is, Val250 and Phe275, could be found in 2vfr. Furthermore, the respective two amino acids are located similarly with respect to the bound ligand. Thus, this double mutation could help to enhance the alditol oxidase activity toward glycerol. Such mutations can have a high impact on neighboring amino acid positions and the stability of the complete structure. Further mutations, such as substituting Pro249 by a smaller amino acid to create more space, could be necessary and detailed modeling experiments should be investigated to verify these suggestions. Nevertheless, the described double mutation and three additional single mutations were experimentally tested by the cooperation partner. Unfortunately, these mutants could not be expressed in the cytoplasm, thus no active oxidases could be produced [69].

This application showed some of the strength and weaknesses of the method. Although valuable information about the active sites of the oxidases and their similarity could be derived, no information about the stability of the enzyme can be added. Thus, in addition to active site comparison and amino acid mutation, molecular dynamic simulations could be applied to computationally analyze the stability of the modified enzyme. Furthermore, computational tools that estimate the stability of protein complexes could be investigated (see Section 5.3.3) to evaluate functionally motivated mutations.

5.4.4 Predicting the Potential of a Target to be Modulated by Low-Molecular-Weight Compounds

Many drug discovery projects in pharmaceutical research fail because the underlying target was afterward found to be undruggable [71]. *A priori* prediction of the potential of a disease-modifying target to be modulated by low-molecular-weight compounds is of major interest to save time and costs in the development pipeline. In this context, the term druggability has been coined in the early 2000s [72] and has been intensively analyzed since then. Similarly, the terms ligandability and targetability are investigated. Thus, the biotechnological question whether an enzyme can be temporarily inhibited by a low-molecular-weight compound could equally be answered by this approach. A prerequisite is only a reasonably sized training set of proteins.

DoGSiteScorer was trained to predict the druggability of protein targets [48]. A prerequisite for the successful application of machine learning techniques, such as SVMs, is a large and reliable data set to train and evaluate the method on. Therefore, a well-characterized druggability data set from literature [73] containing 1069 data points has been chosen. Pockets were detected and descriptors calculated for all targets. Subsequently, a descriptor analysis was performed. In agreement with other literature findings, druggable pockets were found to be larger, more complex, and more hydrophilic than undruggable pockets. In a further step, a model has been trained on a non-redundant version of this data set. Finally, each query structure is annotated as being either druggable or undruggable based on the normalized druggability score returned by the SVM model. The method has been tested on the complete data set, containing bound and unbound structures of proteins from different families, and yielded correct predictions in 88% of the cases. As one important drug target class, kinases that play a major role in cancer and inflammatory diseases are briefly discussed here. The data set contained 40 p38 MAP kinase pockets; almost all of them were correctly assigned as druggable. The underlying structures were crystallized in different activation states (DFG-in and DFG-out), accordingly their volumes ranged from 450 to 1800 Å^3. Nevertheless, other features, such as a high fraction of lipophilic surface area, allowed to classify them as druggable, which exemplifies the robustness of the method to structural changes within the protein structures.

With the aid of this technique, the quality of a new protein of interest can be assessed within seconds and several proteins can be compared and prioritized for biotechnological processes or other investigations based on the calculated scores.

Thus with this technique, novel active and allosteric sites can be explored on proteins for which nothing is known beforehand.

5.4.5 Prediction of Competitive Substrate Inhibition

During the development of synthetic multi-enzyme pathways, wherein the activity of each enzyme should be maximized, competitive substrate inhibition constitutes a bottleneck. In a recently published study, Schomburg *et al*. [74] have established a computational prediction protocol to quantify competitive substrate inhibition by buffering agents. Combining molecular docking using the LeadIT software suite with a rescoring strategy based on the HYDE scoring function, buffering agents that interfere with the binding of the substrate could be identified. In the following, an exemplary case of the effect of buffering agents on the catalytic activity of phosphoglucose isomerase is presented. Further examples of competitive substrate inhibition by buffering agents on enzymatic systems can be found in [74].

Phosphoglucose isomerase catalyzes the reaction of glucose-6-phosphate to fructose-6-phosphate and vice versa (see Figure 5.8). The availability of a 3D structure of phosphoglucose isomerase, preferably co-crystallized with substrate or product, constitutes the basis for studying the effect of buffering agents on its catalytic activity using molecular docking. In the above-mentioned study, a crystal structure of fructose-6-phosphate bound to phosphoglucose isomerase (PDB code: 1hox) was available from the PDB. This structure was prepared within the LeadIT software suite before the docking of 14 different buffering agents was conducted. Afterwards, the binding energy of the proposed molecular complexes – enzyme and a buffering agent – was estimated using the HYDE scoring function and related to the binding affinity of the substrate. This resulted in a relative HYDE score compared with the substrate's HYDE score. Buffering agents exhibiting a relative HYDE score of at least 90% were marked as critical, because they would most likely interfere with the binding of the substrate to the enzyme. Buffering agents whose relative HYDE score is predicted to be in the range of 75–90% also may inhibit the binding of the substrate. Below a relative HYDE score of 75%, the buffering agent is expected not to affect the activity of the enzyme. The computational rating of four buffering agents – carbonate, diglycine, TRIS, and PIPES – on their effect on the activity of phosphoglucose isomerase was afterwards experimentally validated (see Figure 5.8).

To evaluate the predictive power of the relative HYDE score the measured activity of phosphoglucose isomerase in the presence of a buffering agent was converted to a relative activity. This can be done by relating the activity of the enzyme measured in the presence of a buffering agent to the maximum activity the enzyme can achieve with its natural substrate. Using the relative HYDE score PIPES was predicted to highly influence the binding of fructose-6-phosphate to phosphoglucose isomerase. In the experimental test, when PIPES is used as a buffering agent in the activity assay, the measured activity of the enzyme is reduced to 57%.

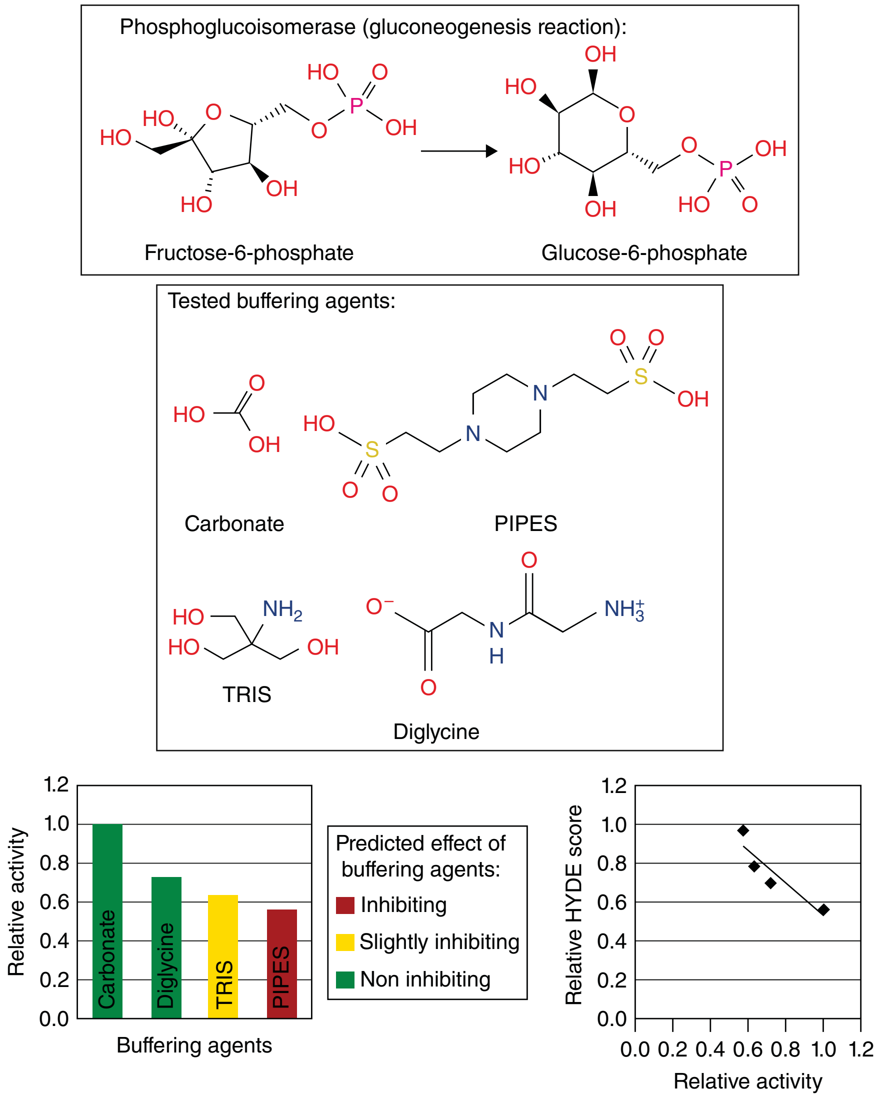


Figure 5.8 Competitive substrate inhibition by buffering agents. (Top) Reaction catalyzed by phosphoglucose isomerase and (Middle) buffering agents tested for interference with the binding of fructose-6-phosphate. (Bottom) Effect on the activity of phosphoglucose isomerase triggered by different buffering agents.

The buffering agent TRIS was also predicted to affect the binding of fructose-6-phosphate to the enzyme. This could be confirmed by a decreased activity of 37% compared with the regular activity of phosphoglucose isomerase. Diglycine as well as carbonate were classified as less critical by the relative HYDE score (diglycine obtaining a relative HYDE score of 70% is not rated as inhibiting). In the activity assay, diglycine reduces the enzyme's activity by 28%. Carbonate, rated as best

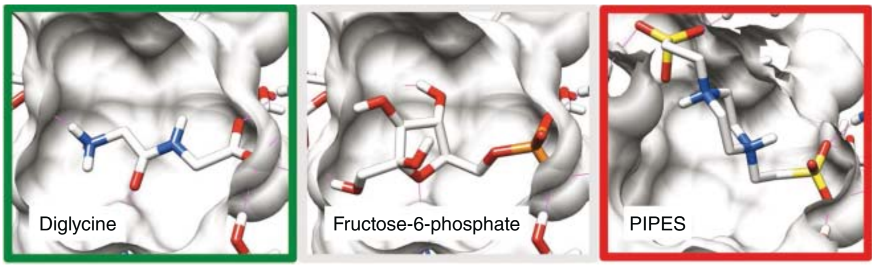


Figure 5.9 Binding modes of phosphoglucose isomerase substrate (fructose-6-phosphate, co-crystallized) (1hox) and two buffering agents, diglycine and PIPES (docking poses). Images were generated with Chimera [75].

buffering agent for phosphoglucose isomerase by the relative HYDE score, has no effect on the binding of fructose-6-phosphate to phosphoglucose isomerase in the experiment.

Figure 5.9 shows the bioactive binding modes of diglycine (left) and PIPES (right) in the phosphoglucose isomerase binding pocket, which were predicted by HYDE. The binding mode of the co-crystallized substrate, fructose-6-phosphate, is also depicted (Figure 5.9, middle). Both buffering agents exhibit a similar binding mode compared with the substrate, partially interacting with the same residue.

5.4.6 Classification of Biological and Artificial Protein Complexes

The interaction of proteins with each other is essential for all biochemical pathways and signal transduction processes. One objective in rational protein design is altering these interactions to analyse the connectivity in signal transduction networks. Protein-protein interactions can be classified according to their lifetime in permanent and transient interactions [76, 77]. Permanent complexes assemble directly after protein transcription and can be characterized by their high stability and long lifetime. In contrast, transient interactions are less stable and thereby reversible, playing a key role in signal transduction processes. Experimental methods to distinguish those protein interactions are time-consuming and exhibit high false-positive rates [78]. An additional type of protein-protein interactions occurs in crystal structures of proteins: the contact of protein subunits artificially forced by the crystallization process. To characterize the function of an enzyme, it is important to known which multimeric state is the natural biological assembly.

In this context, a computational classification method using a machine learning approach and the HYDE scoring function was developed. The basis for the analysis was a data set of 254 protein complexes, which comprises the three above-mentioned types of protein-protein interactions [79], and a set of various descriptors. As a result, it could be revealed that the hydrophobic dehydration energy of

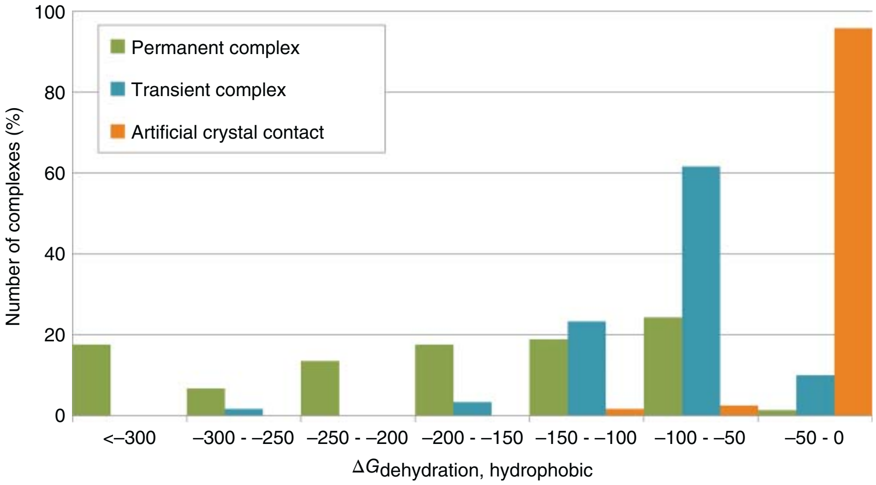


Figure 5.10 Classification of protein-protein contacts using the hydrophobic dehydration energy of the HYDE scoring function. Three different protein-protein contacts were discriminated: permanent and transient complexes, both biologically relevant assemblies, as well as artificial contacts caused due to crystal packing.

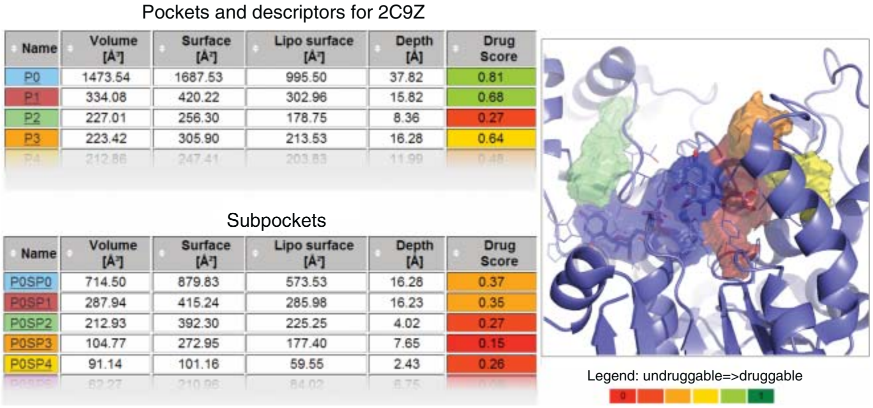
the protein-protein interaction estimated using HYDE is sufficient to classify biologically relevant protein assemblies and artificial crystal complexes with a probability of 95%. Figure 5.10 depicts the distribution of the hydrophobic dehydration energy of the different protein-protein interactions of the data set. Considering the determination of permanent and transient protein interactions, the performance of hydrophobic dehydration energy drops to 73% accuracy (compare Figure 5.10).

This application study shows that the HYDE scoring function can be used to estimate protein-protein interactions and can be applied to automatically classify the biological and artificial protein complexes.

5.4.7
Available Web Services to Support Biocatalysis Research

To provide the functionality of active site prediction and analysis as well as the classification of protein-protein interactions, web services were made publicly available.

The DoGSiteScorer (http://dogsite.zbh.uni-hamburg.de/) [80] offers the full active site analysis functionality in a user-friendly manner. As input, only a 3D protein structure is required (or simply a PDB code). Subsequently, all potential pockets and their properties are calculated on the fly. In Figure 5.11, the results for a flavonoid glycosyltransferase (PDB Code: 2c9z) from red wine



Pockets and descriptors for 2C9Z

Name	Volume [Å³]	Surface [Å²]	Lipo surface [Å²]	Depth [Å]	Drug Score
P0	1473.54	1687.53	995.50	37.82	0.81
P1	334.08	420.22	302.96	15.82	0.68
P2	227.01	256.30	178.75	8.36	0.27
P3	223.42	305.90	213.53	16.28	0.64
P4	212.86	247.41	203.83	11.99	0.48

Subpockets

Name	Volume [Å³]	Surface [Å²]	Lipo surface [Å²]	Depth [Å]	Drug Score
P0SP0	714.50	879.83	573.53	16.28	0.37
P0SP1	287.94	415.24	285.98	16.23	0.35
P0SP2	212.93	392.30	225.25	4.02	0.27
P0SP3	104.77	272.95	177.40	7.65	0.15
P0SP4	91.14	101.16	59.55	2.43	0.26
P0SP5	82.27	210.96	84.02	6.75	[illegible]

Figure 5.11 DoGSiteScorer web server example for a glycosyltransferase (PDB: 2c9z). (Left) Extraction of result tables of (sub)pocket prediction and descriptor calculation. (Right) Subpockets of the largest pocket of 2c9z are shown in different colored isosurfaces. The colors correspond to the subpocket table on the left side. The protein backbone is shown in blue. The co-crystallized ligands (QUE, UDP) are represented in ball-and-stick mode.

with proven *in silico* activity are exemplified. The server outputs a simplified view of the calculated pockets and subpockets together with all pocket properties and a score, which estimates how easy the pockets can be addressed by small compounds. In the case of this flavonoid glycosyltransferase, DoGSiteScorer detects the active sites of the enzyme correctly and furthermore splits the highest ranked pocket into subpockets of which one contains exactly the substrate and another one holds the cofactor. This allows to learn more about the features of the pockets (volume, surface, hydrophobicity, etc.) and to drive further investigations.

The protein-protein interaction classification can also be accessed by a web service (http://ppi.zbh.uni-hamburg.de/). As input, simply a PDB code is required and the user has to specify the protein-protein interface of interest (selection of the corresponding protein chains). Subsequently, the probabilities of being an artificial, permanent, or transient protein-protein interaction are calculated. Additionally to the introduced web services further applications like Protoss can be found on the ProteinsPlus server (http://proteinsplus.zbh.uni-hamburg.de).

5.5 Conclusion and Future Directions

In this chapter, an overview of existing computational tools and methods, which can assist in answering relevant questions in biocatalysis research, has been given.

Novel computational methods have been presented aiming at the discovery and optimization of new catalytic enzymes for biotechnological processes. The basic principles of these methods were presented, and their successful application to various important questions arising during biocatalysis research were exemplified.

A central question in the field of white biotechnology, namely functional analysis of active sites of proteins, was successfully addressed for a comprehensive spectrum of biotechnologically relevant enzymes. It was shown that the developed and optimized approaches for function prediction can be applied to catalytically relevant protein classes as well as to predict enzyme substrate specificity of newly discovered enzymes. The methods can help to gain insight into the catalytic mechanism of an enzyme. Furthermore, they can be used to predict mutation sites in the binding pocket and to optimize the catalytic activity of the enzyme.

The generation of further knowledge about potential binding partners of an enzyme, which can be a particular substrate or another protein, is of special interest in many projects. Scoring functions enable the energetic estimation of protein-ligand as well as protein-protein interactions. In this way, the energetic effect of mutations, the stability analysis of protein structures at different temperatures as well as the improvement of selectivity of protein-protein interactions can be investigated. Furthermore, applying automated alanine scanning – or in more general terms *in silico* mutation of any amino acid residue into any other arbitrary residue – allows a comfortable analysis and optimization of enzymes.

The wide range of possible applications shown in this chapter gives an idea about the usefulness of computational methods for biocatalysis research. The presented computer-based methods constitute only a fraction of computational tools that can and should be intensively used in biotechnological research opening up the wealth of available data and getting new insights and cross-links between structures for rational enzyme design and analysis.

Acknowledgments

We thank all collaboration partners for many fruitful discussions and a successful cooperation. A.V. and M.R. especially thank Thomas Grombacher, Daniel Kuhn, and Friedrich Rippmann of Merck KGaA, Darmstadt. N.S., E.N., and M.R. especially thank Gudrun Lange and Robert Klein of Bayer CropScience AG, Frankfurt. We thank Christian Lemmen and Holger Claussen of BioSolveIT GmbH, Sankt Augustin for their ongoing support during the development of DoGSiteScorer and HYDE. We further thank Sandra Gerstenbruch, Nina Mussmann, and Uwe Bornscheuer for their collaboration in the glycerol oxidase project at Henkel AG. We thank the organizers and the other partners of the Biocatalysis2021 Cluster for being part of this outstanding consortium and enabling effective collaborations. The presented projects were funded by BMBF under grants 0315292A and 031A183B.

References

1. Turner, N.J. (2009) Directed evolution drives the next generation of biokatalysts. *Nat. Chem. Biol.*, **5**, 567–573.
2. Damborsky, J. and Brezovsky, J. (2009) Computational tools for designing and engineering biocatalysts. *Curr. Opin. Chem. Biol.*, **13** (1), 26–34.
3. Damborsky, J. and Brezovsky, J. (2014) Computational tools for designing and engineering enzymes. *Curr. Opin. Chem. Biol.*, **19**, 8–16.
4. Zanghellini, A., Jiang, L., Wollacott, A.M., Cheng, G., Meiler, J., Althoff, E.A., Röthlisberger, D., and Baker, D. (2006) New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.*, **15** (12), 2785–2794.
5. Dahiyat, B.I. and Mayo, S.L. (1996) Protein design automation. *Protein Sci.*, **5** (5), 895–903.
6. Hilvert, D. (2013) Design of protein catalysts. *Annu. Rev. Biochem.*, **82**, 447–470.
7. Kries, H., Blomberg, R., and Hilvert, D. (2013) De novo enzymes by computational design. *Curr. Opin. Chem. Biol.*, **17** (2), 221–228.
8. Malisi, C., Schumann, M., Toussaint, N.C., Kageyama, J., Kohlbacher, O., and Höcker, B. (2012) Binding pocket optimization by computational protein design. *PLoS One*, **7** (12), e52505.
9. Damborsky, J. (1998) Quantitative structure–function and structure–stability relationships of purposely modified proteins. *Protein Eng.*, **11** (1), 21–30.
10. Fox, R.J., Davis, S.C., Mundorff, E.C., Newman, L.M., Gavrilovic, V., Ma, S.K., Chung, L.M., Ching, C., Tam, S., Muley, S., Grate, J., Gruber, J., Whitman, J.C., Sheldon, R.A., and Huisman, G.W. (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.*, **25** (3), 338–344.
11. Petrek, M., Otyepka, M., Banas, P., Kosinova, P., Koca, J., and Damborsky, J. (2006) CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinf.*, **7**, 316.
12. Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sustr, V., Klvana, M., Medek, P., Biedermannova, L., Sochor, J., and Damborsky, J. (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.*, **8** (10), e1002708.
13. Pavlova, M., Klvana, M., Prokop, Z., Chaloupkova, R., Banas, P., Otyepka, M., Wade, R.C., Tsuda, M., Nagata, Y., and Damborsky, J. (2009) Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.*, **5**, 727–733.
14. Tomic, S., Bertosa, B., Wang, T., and Wade, R.C. (2007) COMBINE analysis of the specificity of binding of Ras proteins to their effectors. *Proteins*, **67** (2), 435–447.
15. Massova, I. and Kollman, P.A. (1999) Computational alanine scanning to probe protein–protein interactions: a novel approach to evaluate binding free energies. *J. Am. Chem. Soc.*, **121** (36), 8133–8143.
16. Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2007) Computational alanine scanning mutagenesis—an improved methodological approach. *J. Comput. Chem.*, **28** (3), 644–654.
17. Watanabe, K., Ohkuri, T., Yokobori, S., and Yamagishi, A. (2006) Designing thermostable proteins: ancestral mutants of 3-isopropylmalate dehydrogenase designed by using a phylogenetic tree. *J. Mol. Biol.*, **355** (4), 664–674.
18. Pleiss, J. (2011) Protein design in metabolic engineering and synthetic biology. *Curr. Opin. Biotechnol.*, **22** (5), 611–617.
19. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28** (1), 235–242.
20. Rognan, D. (2011) Docking methods for virtual screening: principles and recent advances, in *Virtual Screening: Principles, Challenges, and Practical Guidelines*, 1st edn (ed C. Sotriffer), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

21. Macchiarulo, A., Nobeli, I., and Thornton, J.M. (2004) Ligand selectivity and competition between enzymes in silico. *Nat. Biotechnol.*, **22** (8), 1039–1045.
22. Kalyanaraman, C., Bernacki, K., and Jacobson, M.P. (2005) Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. *Biochemistry*, **44** (6), 2059–2071.
23. Tyagi, S. and Pleiss, J. (2005) Biochemical profiling in silico – predicting substrate specifities of large enzyme families. *J. Biotechnol.*, **124** (1), 108–116.
24. Seifert, A., Tatzel, S., Schmid, R., and Pleiss, J. (2006) Multiple molecular dynamics simulations of human p450 monooxygenase CYP2C9: the molecular basis of substrate binding and regioselectivity toward warfarin. *Proteins*, **64**, 147–155.
25. Oelschlaeger, P. and Pleiss, J. (2007) Hydroxyl groups in the betabeta sandwich of metallo-beta-lactamases favor enzyme activity: Tyr218 and Ser262 pull down the lid. *J. Mol. Biol.*, **366** (1), 316–329.
26. Hermann, J.C., Marti-Arbona, R., Fedorov, A.A., Fedorov, E., Almo, S.C., Shoichet, B.K., and Raushel, F.M. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature*, **448** (7155), 775–779.
27. Dobson, P. and Doig, A. (2003) Distinguishing enzyme structures from nonenzymes without alignments. *J. Mol. Biol.*, **330** (4), 771–783.
28. Dobson, P. and Doig, A. (2005) Predicting enzyme class from protein structure without alignments. *J. Mol. Biol.*, **345** (1), 187–199.
29. Cai, C., Han, L., Ji, Z., and Chen, Y. (2004) Enzyme family classification by support vector machines. *Proteins*, **55** (1), 66–76.
30. Izrailev, S. and Farnum, M. (2004) Enzyme classification by ligand binding. *Proteins*, **57** (4), 711–724.
31. Bray, T., Doig, A., and Warwicker, J. (2009) Sequence and structural features of enzymes and their active sites by EC class. *J. Mol. Biol.*, **386** (5), 1423–1436.
32. Konc, J. and Janezic, D. (2012) ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. *Nucleic Acids Res.*, **40** (Web Server issue), W214–W221.
33. Sotriffer, C. and Matter, H. (2011) The challenge of affinity prediction: scoring functions for structure-based virtual screening, in *Virtual Screening: Principles, Challenges, and Practical Guidelines*, 1st edn (ed C. Sotriffer), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.
34. Janin, J. and Rodier, F. (1995) Protein–protein interaction at crystal contacts. *Proteins*, **23** (4), 580–587.
35. Levy, E.D. (2007) PiQSi: protein quaternary structure investigation. *Structure*, **15** (11), 1364–1367.
36. Mitra, P. and Pal, D. (2011) Combining Bayes classification and point group symmetry under Boolean framework for enhanced protein quaternary structure inference. *Structure*, **19** (3), 304–312.
37. Leach, A.R. (2001) *Molecular Modelling: Principles and Applications*, Pearson Education.
38. Gu, J. and Bourne, P.E. (eds) (2009) *Structural Bioinformatics*, vol. **44**, John Wiley & Sons, Inc.
39. Mannhold, R., Kubinyi, H., and Folkers, G. (2011) in *Virtual Screening: Principles, Challenges, and Practical Guidelines*, vol. **48** (ed C. Sotriffer), John Wiley & Sons, Inc.
40. LeadIT. BioSolveIT GmbH, Sankt Augustin http://www.biosolveit.de/leadit/ (accessed 26 February 2014).
41. Reulecke, I., Lange, G., Albrecht, J., Klein, R., and Rarey, M. (2008) Towards an integrated description of hydrogen bonding and dehydration: reducing false positives in virtual screening with the HYDE scoring function. *ChemMedChem*, **3** (6), 885–897.
42. Schneider, N., Lange, G., Hindle, S., Klein, R., and Rarey, R. (2013) A consistent description of HYdrogen bond and DEhydration energies in protein–ligand complexes: methods behind the HYDE scoring function. *J. Comput. Aided Mol. Des.*, **27** (1), 15–29.
43. Fischer, E. (1894) Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges. Berlin*, **27** (3), 2985–2993.

44. Gold, N. and Jackson, R. (2006) A searchable database for comparing protein-ligand binding sites for the analysis of structure-function relationships. *J. Chem. Inf. Model.*, **46** (2), 736–742.

45. Volkamer, A. and Rarey, M. (2014) Exploiting structural information for drug- target assessment. *Future Med. Chem.*, **6**(3), 319–331.

46. Volkamer, A., Griewel, A., Grombacher, T., and Rarey, M. (2010) Analyzing the topology of active sites: on the prediction of pockets and sub-pockets. *J. Chem. Inf. Model.*, **50** (11), 2041–2052.

47. Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2** (3), 1–27.

48. Volkamer, A., Kuhn, D., Grombacher, T., Rippmann, F., and Rarey, M. (2012) Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.*, **52** (2), 360–372.

49. Volkamer, A., Kuhn, D., Rippmann, F., and Rarey, M. (2013) Predicting enzymatic function from global binding site descriptors. *Proteins*, **81** (3), 479–489.

50. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, **261** (3), 470–489.

51. Koshland, D.E. (1958) Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.*, **44** (2), 98–104.

52. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E. (1982) A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, **161** (2), 269–288.

53. Jones, G., Willett, P., and Glen, R.C. (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.*, **245** (1), 43–53.

54. Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Daniel, T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., Shaw, D.E., Francis, P., and Shenkin, P.S. (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, **47** (7), 1739–1749.

55. Claußen, H., Buning, C., Rarey, M., and Lengauer, T. (2001) FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, **308** (2), 377–395.

56. Corbeil, C.R., Englebienne, P., and Moitessier, N. (2007) Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.*, **47** (2), 435–449.

57. Lippert, T. and Rarey, M. (2009) Fast automated placement of polar hydrogen atoms in protein-ligand complexes. *J. Cheminf.*, **1** (1), 1–12.

58. Bietz, S., Urbaczek, S., Schulz, B., Rarey, M. (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *Journal of Cheminformatics*, **6** (12), 1–12.

59. Schneider, N., Hindle, S., Lange, G., Klein, R., Albrecht, J., Briem, H., Beyer, K., Claußen, H., Gastreich, M., Lemmen, C., and Rarey, R. (2012) Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *J. Comput. Aided Mol. Des.*, **26** (6), 701–723.

60. Lange, G., Klein, R., Albrecht, J., Rarey, M., and Reulecke, I. (2010) Method for the determination of intra- and intermolecular interactions in aqueous solution. European Patent EP2084520 B1, filed Oct. 20, 2007 and issued Sep. 15, 2010.

61. Schneider, N., Klein, R., Lange, G., and Rarey, M. (2012) Nearly no scoring function without a Hansch-analysis. *Mol. Inf.*, **31**, 503–507.

62. Hubbard, T., Murzin, A., Brenner, S., and Chothia, C. (1997) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **25** (1), 236–239.

63. Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., and Thornton, J. (1997) CATH – a hierarchic classification of protein domain structures. *Structure*, **5** (8), 1093–1109.

64. Krissinel, E. and Henrick, K. (2014) Protein Structure Comparison Service

Fold at European Bioinformatics Institute, http://www.ebi.ac.uk/msd-srv/ssm (accessed 26 February 2014).

65. Taubig, H., Buchner, A., and Griebsch, J. (2006) PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.*, **34** (Web Server issue), W20–W23.

66. Gibrat, J., Madej, T., and Bryant, S. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6** (3), 377–385.

67. Nadzirin, N. and Firdaus-Raih, M. (2012) Proteins of unknown function in the Protein Data Bank (PDB): an inventory of true uncharacterized proteins and computational tools for their analysis. *Int. J. Mol. Sci.*, **13** (10), 12761–12772.

68. Institut für Technische Biochemie (2014) Enzyme Database, Processed in the Group of Prof. Pleiss, www.itb.uni-stuttgart.de/research/bioinformatics (accessed 26 February 2014).

69. Gerstenbruch, S., Wulf, H., Mussmann, N., O'Connell, T., Maurer, K.H., and Bornscheuer, U.T. (2012) Asymmetric synthesis of D-glyceric acid by an alditol oxidase and directed evolution for enhanced oxidative activity towards glycerol. *Appl. Microbiol. Biotechnol.*, **96** (5), 1243–1252.

70. Chemical Computing Group Inc. (2014) Molecular Operating Environment (MOE), www.chemcomp.com/MOE-Molecular_Operating_Environment.htm (accessed 26 February 2014).

71. Brown, D. and Superti-Furga, G. (2003) Rediscovering the sweet spot in drug discovery. *Drug Discovery Today*, **8** (23), 1067–1077.

72. Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discovery*, **1** (9), 727–730.

73. Schmidtke, P. and Barril, X. (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.*, **53** (15), 5858–5867.

74. Schomburg, K.T., Ardao, I., Götz, K., Rieckenberg, F., Liese, A., Zeng, A.-Z., and Rarey, M. (2012) Computational biotechnology: prediction of competitive substrate inhibition of enzymes by buffer compounds with protein-ligand docking. *J. Biotechnol.*, **161** (4), 391–401.

75. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004) UCSF Chimera visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25** (13), 1605–1612.

76. Jones, S. and Thornton, J.M. (1996) Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.*, **93** (1), 13–20.

77. Tsai, C.J. and Nussinov, R. (1997) Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association. *Protein Sci.*, **6** (7), 1426–1437.

78. Deane, C.M., Salwiński, Ł., Xenarios, I., and Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, **1** (5), 349–356.

79. Zhu, H., Domingues, F., Sommer, I., and Lengauer, T. (2006) NOXclass: prediction of protein-protein interaction types. *BMC Bioinf.*, **7**, 27.

80. Volkamer, A., Kuhn, D., Rippmann, F., and Rarey, M. (2012) DoGSiteScorer: a web-server for automatic binding site prediction, analysis, and druggability assessment. *Bioinformatics*, **28** (15), 2074–2075.

Evidence of Water Molecules – A Statistical Evaluation of Water Molecules Based on Electron Density.

[D2] **Nittinger, E.**; Schneider, N.; Lange, G.; Rarey, M. Evidence of Water Molecules – A Statistical Evaluation of Water Molecules Based on Electron Density. J. Chem. Inf. Model. 2015, 55 (4): 771–783.

http://pubs.acs.org/articlesonrequest/AOR-EaZzEVcctaj8JPManqyp

Journal of
CHEMICAL INFORMATION
AND MODELING

Article

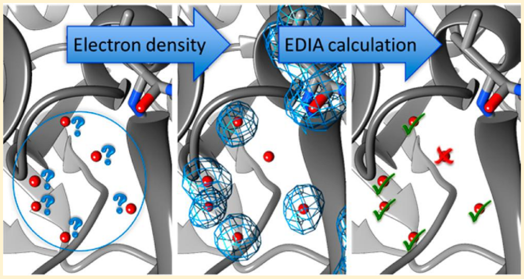
pubs.acs.org/jcim

Evidence of Water Molecules—A Statistical Evaluation of Water Molecules Based on Electron Density

Eva Nittinger,[†] Nadine Schneider,[†] Gudrun Lange,[‡] and Matthias Rarey*,[†]

[†]Center for Bioinformatics, University of Hamburg, Bundesstraße 43, 20146 Hamburg, Germany
[‡]Bayer CropScience AG, Industriepark Hoechst, G836, 65926 Frankfurt am Main, Germany

Supporting Information



ABSTRACT: Water molecules play important roles in many biological processes, especially when mediating protein−ligand interactions. Dehydration and the hydrophobic effect are of central importance for estimating binding affinities. Due to the specific geometric characteristics of hydrogen bond functions of water molecules, meaning two acceptor and two donor functions in a tetrahedral arrangement, they have to be modeled accurately. Despite many attempts in the past years, accurate prediction of water molecules—structurally as well as energetically—remains a grand challenge. One reason is certainly the lack of experimental data, since energetic contributions of water molecules can only be measured indirectly. However, on the structural side, the electron density clearly shows the positions of stable water molecules. This information has the potential to improve models on water structure and energy in proteins and protein interfaces. On the basis of a high-resolution subset of the Protein Data Bank, we have conducted an extensive statistical analysis of 2.3 million water molecules, discriminating those water molecules that are well resolved and those without much evidence of electron density. In order to perform this classification, we introduce a new measurement of electron density around an individual atom enabling the automatic quantification of experimental support. On the basis of this measurement, we present an analysis of water molecules with a detailed profile of geometric and structural features. This data, which is freely available, can be applied to not only modeling and validation of new water models in structural biology but also in molecular design.

■ INTRODUCTION

In order to fully understand complex biomolecular structures, the role of water molecules needs to be comprehended in greater detail. Water molecules form part of the environment of biological macromolecules, in which they can on the one hand stabilize protein folding by mediating interactions and on the other sustain the dynamics of the protein.[1−5] Enzymatic reactions often directly involve one water molecule, i.e., as reactant for hydrolysis reactions,[6,7] or as steric hindrance to guide stereoselectivity[8]. Further, water molecules stabilize biological complexes by mediating protein−protein or protein−ligand interactions, in which mediated hydrogen bonds are often as abundant as direct interactions.[9−12]

In addition to the aforementioned biological processes, water molecules also play an essential role in energetic effects upon binding of, for example protein and ligand or protein and protein, due to their contribution to hydration and dehydration as well as the hydrophobic effect.[13−19] On the one hand, energy is needed in order to dehydrate hydrophilic atoms of the protein and the ligand. On the other hand, energy is also gained by releasing water molecules into the bulk solvent and the hydrophobic effect. Therefore, in order to correctly estimate protein−ligand binding affinities, water molecules have to be taken into account when developing drugs. However, it is not yet understood how water molecules exactly contribute to the binding affinity.[20,21] It is hardly possible to experimentally measure the energy contribution of an individual water molecule. Even if a water molecule in the binding site can be displaced, for instance by an extension of the ligand, the resulting binding affinity is a combination of those two effects, replacement and substitution.[22−24]

Different computational methods have been developed lately that try estimate the energetic cost or gain of water molecule displacement to guide rational drug design. Herein, two different approaches exist: first, classification of X-ray crystallographic water molecules[25−28] and, second, computer based prediction of water molecule positions[29−36] (Table 1). As early as 1985, Goodford developed a method, which calculates energies between diverse probe groups and a protein in a grid-based manner.[30] One of these probes resembled water, enabling the prediction of water molecule positions in three examples in good agreement to X-ray crystallographic water molecules. However, its overall applicability has not been proven.

Received: October 31, 2014
Published: March 5, 2015

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

Table 1. Methods for Positioning Water Molecules within Biological Complexes and Prediction of Water Molecule Stability—Structurally and Energetically

method type	method name	prediction task
simulation-based	Grand Canonical Monte Carlo simulation (GCMC)[29]	binding free energy estimation
	GIST[35]	displaceability, thermodynamic analysis
	JAWS[34]	binding affinity estimation
	SZMAP[31]	orientation and displaceability
	WaterMap[32,33]	energy (enthalpic and entropic) contribution
docking-based	WaterDock[36]	conserved vs displaceable
grid-based	GRID[30]	energy (enthalpic) contribution
	WaterFlap	water score ("happiness")

Recently, more elaborate and time-expensive methods have been developed, from which some use molecular dynamics simulations in order to place water molecules and estimate their binding affinity contributions (e.g., WaterMap[32,33]). These programs can be separated into three main types−simulation-based, docking-based, and grid-based. Furthermore, they can be classified according to their overall prediction aim (See Table 1). Most of the time it is not known which water molecules will be displaced upon binding and which ones remain in the protein binding site, thus mediating between protein and ligand. However, it is assumed that those water molecules remaining in the protein complex are more stable than the ones being displaced upon complex formation.

Diverse characterizations of water molecules have been conducted, ranging from structural analysis to thermodynamic description.[37−44] Water molecules in protein−ligand interactions have been of great interest and different analyses have shown that those bridging protein and ligand often have three or more interactions compared to only one interaction on average in protein−protein interfaces.[37,38] Furthermore, water molecules prefer interactions to the backbone rather than to the side chain, indicated by the number of interactions as well as thermodynamically.[38−40] Dunitz has approximated the entropic gain of transferring a water molecule from protein into bulk water to be up to 7 cal/mol at room temperature.[41] Using inhomogeneous fluid solvation theory the impact of α-helix and β-sheet on the thermodynamic properties of water molecules has been analyzed. Herein, the thermodynamics of water molecules are affected up to a distance of 4 Å from β-sheets and 4.3 Å from α-helices.[42,43] In an application of the method WaterMap, it was estimated that charged amino acids display the most favorable hydration sites, whereas backbone, aromatic, and aliphatic amino acids are less favorable.[44] In this publication we want to keep the focus on structural characteristics of crystallographic water molecules. More detailed information about thermodynamics of water molecules can be found in ref 45.

In order to retrieve experimental data for water molecules in biological complexes, crystal structures are the major source. However, X-ray crystallographic experiments result in diffraction patterns, which have to be interpreted further in order to acquire the molecular structure. The temperature factor (B-factor) is an indicator of thermal motion of each atom and is often taken as a criterion to identify flexible regions in protein structures. However, the B-factor depends on the refinement procedure: its interpretation can be artifactural if crystal contacts are neglected and it varies between different structures. The B-factor does not inform whether an atom is resolved by electron density, but it does indicate its structural flexibility and disorder. Electron density, which is available for many structures nowadays, is the fundamental experimental data available for water molecules (See Table 2). Two measurements exist that describe electron density for specific parts of a structure: the real-space R-factor (RSR) and the real-space correlation coefficient (RSCC).[46] The RSR was developed as an objective interpretation of electron density maps and for the localization of errors during density map interpretation. It is calculated by using the observed electron density from the crystallographic experiment and the calculated electron density derived from the built structure. The lower the RSR is (in a range from 0 to 1) the better the fit of the structure in the electron density. The RSCC, in contrast to the RSR, is the correlation coefficient between the two maps resulting in a value between −1 (complete anticorrelation) and 1 (complete correlation). One drawback of the RSCC arises when it is calculated for atoms with weak densities, but correct intensity distributions. This is especially problematic for water molecules since even a good score might be achieved with low resolution. For protein−ligand complexes, it was shown lately by Hawkins et al.[47] that neither RSR nor RSCC can adequately capture the difference between observed and calculated data.

Table 2. Values Used as Structural Quality Criteria for Identification of Modeling Errors and Structural Uncertainties

value	advantage	disadvantage
B-factor	• included in PDB file • indicator of thermal motion	• can be misinterpreted, e.g. due to crystal packing
RSR	• normalized value [0 (good) to 1 (bad)]	• resolution and density threshold dependent
RSCC	• no density threshold used • normalized value [−1 (anticorrelation) to 1 (correlation)]	• weak density with correct intensity distribution might lead to a good score

Water molecules resolved by X-ray crystallography exist at local energy minima of the position.[48] However, not all atoms present in molecular structures are supported by electron density. Most water molecules are too flexible to be resolved. Furthermore, water molecules found in the crystallographic structure vary strongly in their experimental support. For a detailed study of water molecules in proteins and at protein interfaces, a quantitative measure of electron density support is therefore mandatory to exclude noise arising from unresolved water molecules.

In order to exploit electron density for this task, we developed a new value, called EDIA (= **E**lectron **D**ensity for **I**ndividual **A**toms), which describes the experimental electron density around a single atom, for instance around a single water molecule. Using EDIA, we conducted an extensive evaluation of water molecules from a high-resolution subset of the PDB[49], containing 5485 PDB structures and over 2.3 million water molecules. In the following sections, we give a detailed description of the selected data set and its diversity. The newly developed EDIA measure will be explained in detail. According to the EDIA value, the water molecules of the data set are examined with regard to several structural and geometric characteristics, such as hydrogen bonding preferences and their structural environment. Finally, we show detailed examples of frequently discussed issues like "hydrophobic bubbles"[40] or modeling errors, which can still be found in high-resolution data.

■ MATERIALS AND METHODS

Data Set. A high-resolution Protein Data Bank[49] subset was compiled using the following advanced search criteria: resolution ≤ 1.5 Å, experimental method = X-ray, molecule type = protein. All structures between 2000-01-01 and 2014-02-01 were selected that had an external link to the EDS server[50], ensuring the availability of electron density data for all chosen structures. Using all criteria, a data set of 5526 PDB structures was compiled (Figure 1, date of download: February 1, 2014).

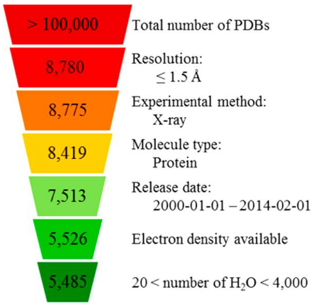


Figure 1. Data set compilation and the effect of each search criterion on the number of PDB structures.

In the next step, two extremes were discarded: those PDB structures with less than 20 water molecules and those with more than 4000, resulting in a final data set of 5485 PDB structures with 2 330 581 water molecules (See Table S1).

Electron Density-Based Value. The electron density is provided as a 3D grid for the asymmetric unit, the smallest unit of volume that by application of symmetry operations is able to reconstruct the unit cell. The unit cell on the other hand is the smallest volume that only by translational application can recreate its pattern in space. We developed an automated estimation of electron density around a single atom not covalently bound to other heavy atoms, called Electron Density for Individual Atoms (EDIA). EDIA is the weighted sum of experimental electron density values around a single atom a in its van der Waals radius:

$$\mathrm{EDIA}(a) = \frac{1}{\sum_{p \in S(a)} \omega(p) \cdot \sigma} \sum_{p \in S(a)} \omega(p) \cdot (f(p) - \mu) \tag{1}$$

where $\omega(p)$ describes a weight function for grid point p, σ, the electron density threshold, and $S(a)$ the subset of grid points around an atom a. The function $f(p)$ reflects the density value at grid point p, and μ, the mean density of the electron density map. In order to allow comparisons between different structures, the EDIA is normalized by the standard deviation σ of the respective electron density map of the asymmetric unit.

Each grid point of the electron density map is associated with a measured density value. In eq 1 $S(a)$ is the subset of all grid points G that are within the van der Waals radius of atom a:

$$S(a) = \{p \in G \,|\, |\overrightarrow{px_a}| \leq r_{\mathrm{vdW}(a)}\} \tag{2}$$

The distance in angstroms of a grid point p to the atom center x_a is $|\overrightarrow{px_a}|$. The distribution of electron density around an atom, caused by the vibration of the atom itself as well as the distribution of electrons around an atom, is resembled using a Gaussian weight:

$$\omega(p) = \mathrm{e}^{-1/2(|\overrightarrow{px_a}|/\delta)^2} \tag{3}$$

The width of the Gaussian bell δ was defined as the covalent radius of the atom. This results in a weight of 0.5 when the distance of a grid point to the center of the atom is equal to its covalent radius. The Gaussian distribution was combined with a linear function $g(p)$ in order to get no further density contributions of grid points with a distance greater than the van der Waals radius of the atom (Figure 2a):

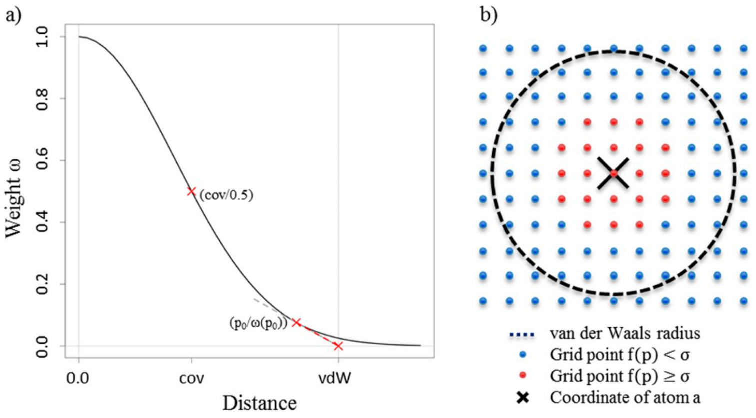


Figure 2. (a) Gaussian weight combined with linear function for EDIA calculation. (b) 2D scheme of an atom and its surrounding electron density grid, with grid points contributing (red dots) and not contributing (blue dots) to EDIA: cov = covalent radius of atom a, vdW = van der Waals radius of atom a, $(p_0/\omega(p_0))$ = starting point of linear function (red dashed line).

$$g(p) = \frac{\omega(p_0)}{p_0 - r_{vdW}}(p - r_{vdW}) \tag{4}$$

The slope of the linear function was selected such that the overall function remains continuously differentiable by passing through the points $(p_0, \omega(p_0))$ and $(r_{vdW}, 0.0)$, with $p_0 = (r_{vdW}/2) + [(r_{vdW}^2/4) - \delta^2]^{1/2}$.

Only density values ρ at grid points p that are above the density threshold σ (= map mean density + standard deviation, threshold typically applied for density visualization) were added up (Figure 2b):

$$f(p) = \begin{cases} \rho(p), & \rho(p) \geq \sigma \\ 0, & \rho(p) < \sigma \end{cases} \tag{5}$$

By excluding very low density values, the noise in the EDIA can be substantially reduced.

Preprocessing of PDB Structures. In order to analyze the hydrogen bond network, hydrogen positions were calculated using Protoss.[51] Protoss adds hydrogens and places them into the structure optimizing the hydrogen bond network accounting for tautomers and protonation states. Herein, it not only considers the amino acids of the protein but also water molecules, metals and ligands. Using an empirical scoring function, Protoss generates an optimal hydrogen bond network for a biological complex. This preprocessing step was applied once for each PDB complex. Afterward, each water molecule of the complex was analyzed individually in its surrounding (4.5 Å radius around the center of the water oxygen).

Descriptor Calculation. For the characterization of water molecules within their surrounding environment, several descriptors were calculated (See Table 3).

Table 3. Summary of Descriptors Used to Characterize Water Molecules

descriptor type	description	details
hydrogen bond descriptors	number of hydrogen bonds	• to all modeled atoms • to protein and ligand atoms as well as metals
	mean length of hydrogen bonds	• to all modeled atoms • to protein and ligand atoms as well as metals
	hydrogen bond partners	• atom type • functional group • acceptor or donor • side chain or backbone
proximity-based descriptors	hydrophobicity	• proportion of hydrophobic atoms within 4.5 Å radius
	hydrophobic surface	• proportion of hydrophobic surface pointing toward the water molecule
	water clusters	• water molecules within 3.5 Å radius

Hydrogen Bond Descriptors. Hydrogen bonds of water molecules to donors or acceptors were identified if the opening angle between ideal donor and acceptor direction was less than 50° and the distance between hydrogen and acceptor was within a range of 1.9 ± 0.5 Å. For each hydrogen bonding function only the geometrically best hydrogen bond was accepted. In this way, bifurcate hydrogen bonds were excluded. The total number of hydrogen bonds and the mean length of the hydrogen bonds were calculated for each water molecule. Additionally, the hydrogen bonding partners were recorded, i.e., atom type, functional group, acceptor or donor, backbone or side chain.

Hydrophobicity-Based Descriptors. In order to classify the surrounding of a water molecule, two types of hydrophobicity-related values were applied. First, the fraction of hydrophobic atoms in a 4.5 Å surrounding sphere was calculated (Figure 3a):

$$\text{hydrophobicity} = \frac{\begin{array}{c}\text{number of hydrophobic atoms}\\ \text{surrounding a water molecule (4.5 Å)}\end{array}}{\begin{array}{c}\text{total number of atoms}\\ \text{surrounding a water molecule (4.5 Å)}\end{array}} \tag{6}$$

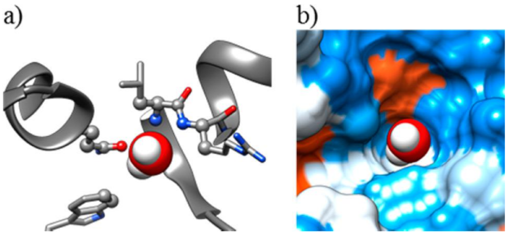


Figure 3. Hydrophobicity-based descriptors. (a) Ball-and-stick represented atoms are within 4.5 Å distance and contribute to hydrophobicity (hydrophobicity = 0.623). (b) Molecular surface of a water molecule surrounded by protein (hydrophobic surface = 0.372): orange = hydrophilic, blue = hydrophobic (molecular graphics were created using UCSF Chimera[52]).

Second, the size of the hydrophobic surface patches of surrounding atoms (≤4.5 Å) was calculated. Here, only those surface patches being closer than 2 Å to the water molecule were considered. For normalization, the fraction was used. (Figure 3b):

$$\text{hydrophobic surface} = \frac{\begin{array}{c}\text{surface area of hydrophobic atoms}\\ \text{pointing towards a water molecule (2 Å)}\end{array}}{\begin{array}{c}\text{surface area of all atoms}\\ \text{pointing towards a water molecule (2 Å)}\end{array}} \tag{7}$$

Water Cluster. The water content of the surroundings of a water molecule was analyzed using water clusters. Herein, within a distance of 3.5 Å, the surrounding was checked for other water molecules. If another water molecule is present, another test, which checked for the presence of further water molecules, was performed. This procedure was prolonged until no further water molecule was identified. Finally, the total number of water molecules within one cluster was counted.

Allocation of Water Molecules. Since water molecules occupy different positions within a biological complex, a further classification into surface (S), protein–ligand interface (PLI), protein–protein interface (PPI), and captured (C, also often called "buried") was performed. This categorization was carried out using the molecular surface area (MSA, Figure 4). Water molecules were classified as PLI, if protein and ligand atoms were found within a 4.5 Å radius around the oxygen atom. Analogous classification was performed for PPI water molecules, which have atoms of two different protein chains

DOI: 10.1021/ci500662d
J. Chem. Inf. Model. 2015, 55, 771−783

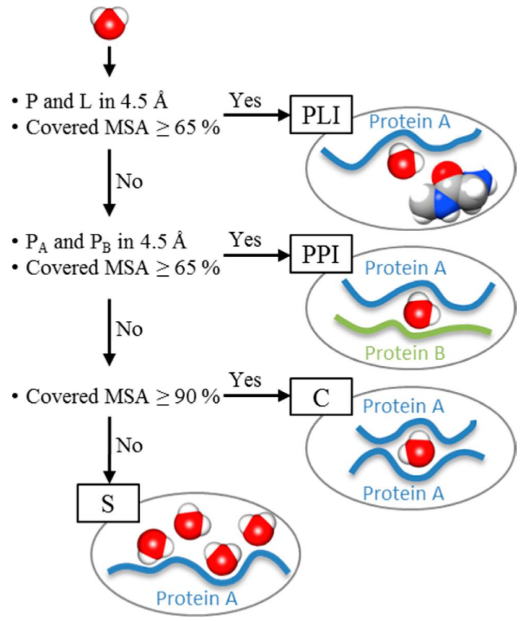


Figure 4. Classification of water molecules: P = protein, L = ligand, P_A = protein A, P_B = protein B.

within a radius of 4.5 Å. In both cases, we additionally checked that at least 65% of their MSA was covered, since they would otherwise lie in the outer rim of either PLI or PPI. Water molecules were classified as C if more than 90% of their MSA was covered by a single protein chain, which means it has more than three covered hydrogen bonding functions. All remaining water molecules were classified as S.

Data Set Compositions. The high-resolution data set was divided into different subsets to allow a detailed structural analysis of different classes of water molecules. Apart from the classification into S, C, PLI, and PPI, the data set was further classified according to diverse structural criteria (See Table 4).

RESULTS AND DISCUSSION

Data Set Composition. The data set is highly diverse with respect to the water content of the PDB structures (Figure 5a). Some complexes have very low water content with less than one water molecule per two amino acids (13%). Others in contrast are more highly hydrated with more than three water molecules per two side chains (19%). One of the "dry" proteins is a heat-labile enterotoxin of *E. coli* with one water molecule per 100 amino acids (PDB ID: 4fo2). Its biological unit is a pentamer; mostly water molecules of the inner protein parts were modeled while an outer solvation layer is nearly absent.

Table 4. Subsets of the High Resolution Data Set, Their Compositions, and Abbreviations Used for the Analysis of Water Molecules

data set abbreviation	data set composition
$\text{Hbond}_{\text{all}}$	water molecules interacting with protein, ligand, and water molecules
Hbond_{PL}	water molecules interacting with protein and ligand
$\text{Hbond}_{\text{H2O}}$	water molecules interacting with water molecules
HB	water molecules that cannot form any hydrogen bond and lie within a highly hydrophobic surrounding (= hydrophobic bubbles)
WIND_{all}	well-integrated water molecules (≥three hydrogen bonds) without electron density interacting with protein, ligand, and water molecules
WIND_{PL}	well-integrated water molecules (≥three hydrogen bonds) without electron density interacting with protein or ligand

A structure of a "wet" protein is an antifreeze protein of *L. dearborni*, which consists of only one very small subunit (63 amino acids) and up to four water molecules per amino acid (PDB ID: 1ucs). This protein does not have any inner water molecules (C, PLI, or PPI), but an extensive solvation layer with many close water molecules. The structure is very well-resolved (0.62 Å) and even hydrogen atoms could be modeled. However, the distances between water molecules are often very small, if not even clashing. The intention of the authors was to model different interaction networks, which are indicated by reduced occupancies of water molecules.[53] Multiple water molecules are modeled into the electron density of oxygen and hydrogen atoms (Figure 6a). This example (PDB ID: 1ucs) shows that even if the structures are of high-resolution, it should not be taken as a guarantee for easily interpretable data. Few other modeling errors also captured our interest, such as the fusion of a water molecule with an amino acid side chain (Figure 6b). This error would not be detected with EDIA, because electron density from the amino acid is available. However, it can be detected using either the electron density difference map ($f_o - f_c$ map), since too many electrons are available; or the difference of Gaussian operator, since the position of the water molecule has clearly no circular density distribution. Overlaps of water molecules were also identified in 11 PDB structures. Those contain multiple water molecules with identical coordinates of the oxygen atom (PDB ids: 2ghc, 2hc1, 2yqb, 3t6f, 3ziy, 3zjp, 4a8n, 4af9, 4ayp, 4ayr, 4b8x). Furthermore, two structures included overlaid ligands with different occupancies (overlap of biotin and biotin-D-sulfoxide (PDB id: 3t6f), overlay of *β*-D-glucose and *α*-D-glucose (PDB id: 4af9), in both cases the ligand with the higher occupancy was kept for further analysis) and 36 structures contained unknown ligands that were represented by oxygen atoms only (See Table S1).

In literature it is often mentioned that the number of observed water molecules depends on the resolution of the structure.[10,48,54,55] Figure 5b shows that within structures of our high-resolution data set, the ratio of water molecules hardly varies. In all cases, the median number of water molecules per amino acid is close to one, with a minimum of one water molecule per 100 amino acids (PDB id: 4fo2) and a maximum of nearly four water molecules per one amino acid (PDB id: 1ucs).

The data set compromises a wide range of protein complexes including complexes without any ligands (25%) and those containing ligands, cofactors, and crystallization buffer molecules (75%). Protein complexes range from small monomers to multimers, with more than 40% of the data set containing more than one protein chain. The size of the proteins ranges from 12 (peptides) to 5480 amino acids. Using the classification of water molecules described in the Materials and Methods section, the data set contains 127 988 PPI and 80 717 PLI water molecules, 264 584 captured (C), and 1 857 292 surface (S) water molecules (Figure 7).

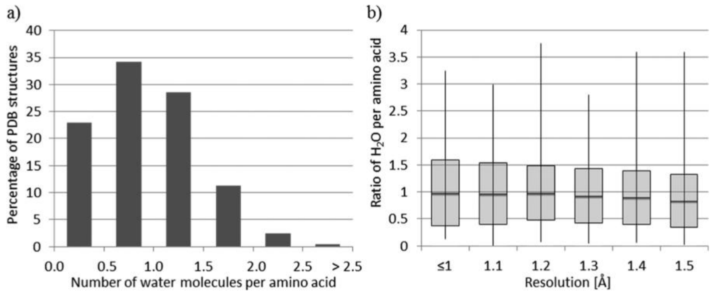


Figure 5. (a) Histogram of the number of water molecules per amino acid for the data set of 5485 PDB structures: median = 0.89, standard deviation = 0.51. (b) Box plot of resolution dependent number of water molecules per amino acid: number of water molecules per category (≤1 Å) 177, (1.1 Å) 296, (1.2 Å) 487, (1.3 Å) 826, (1.4 Å) 1349, (1 Å) 2340. Box limits are median ± one standard deviation.

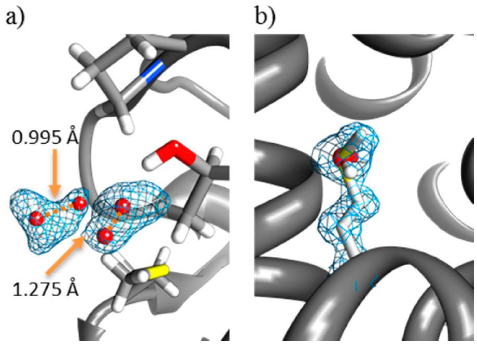


Figure 6. (a) Two alternative water molecules are modeled into the density of one water molecule, H_2O-A-256—H_2O-A-182: 1.275 Å, H_2O-A-284—H_2O-A-136: 0.995 Å (PDB id: 1ucs). (b) Oxygen atom of the water molecule B-2179 fuses with sulfur atom of methionine B-208 (PDB id: 2jae). Electron density is only shown for methionine and water molecule. Difference electron density (not displayed) indicates too many electrons at the water position: blue = electron density map ($2f_o - f_c$) at 1σ (molecular graphics were created using UCSF Chimera[52]).

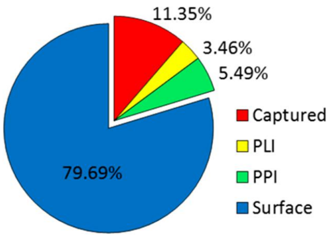


Figure 7. Classification of water molecules according to their position in the complex and their contribution to the data set of 2.3 million water molecules is displayed.

EDIA Values. The histogram of all EDIA values for the high-resolution data set approximately displays an extreme value distribution (Figure 8a). The individual EDIA values resemble well the graphically observable electron density (Figure 8b–f), with zero meaning no electron density and values above one describing clear electron density. Visual inspection of water molecules, their corresponding electron density, and the EDIA give confidence that the EDIA captures the measured electron density.

EDIA was further used to differentiate between well-resolved (with clear electron density) and insufficiently resolved (insufficient electron density) water molecules. As a cutoff value we used $EDIA_{Thrs}$ = 0.24, which is the median EDIA value of all water molecules from the data set minus one standard deviation. The median was chosen, because it is less affected by outliers, especially arising from a few very high EDIA values. Note that a cutoff value of zero was not used because surrounding and very close atoms might cause a slight increase of a water's EDIA value. While this has very little effect on resolved water molecules, it leads to a very small EDIA for unresolved water molecules. Visual inspection confirmed the chosen threshold of 0.24 (see Figure 8c and d), above which water molecules are considered as sufficiently resolved. This leads to 8.9% (208 052) of all water molecules of the data set being classified as insufficiently resolved by electron density, from which the majority (93.4%) belongs to the group of surface water molecules (See Table 5). This meets previous expectations and confirms the applicability of the EDIA.

Hydrogen Bonding Characteristics of Water Molecules. Water molecules were analyzed for their hydrogen bonding characteristics, wherein a maximum of four hydrogen bonds, two donor and two acceptor functions, was assumed. Bifurcated hydrogen bonds were excluded by considering only the geometrically best hydrogen bond for each hydrogen bonding function (see the Materials and Methods section).

Three separate statistics were created: (1) all hydrogen bonds to explicitly modeled atoms were counted for all water molecules, (2) only water molecules sufficiently resolved by electron density ($EDIA \geq EDIA_{Thrs}$) were considered, (3) only water molecules unresolved by electron density ($EDIA < EDIA_{Thrs}$) were taken into account (see Table 6).

The first statistic, which includes all water molecules, results in a mean number of hydrogen bonds of 2.15 ($Hbond_{total}$) from which 1.22 ($Hbond_{H2O}$) are formed to other water molecules. As expected, the second statistic, describing only water molecules with clear electron density, does not vary much from the first one. However, looking at water molecules without clear electron density in the third statistic, the mean number of hydrogen bonds decreases significantly ($Hbond_{total}$ = 1.65). In both sets, meaning water molecules resolved by electron density and unresolved ones, the high proportion of surface water molecules causes a bias in the mean number of hydrogen bonds (See Table 5 and Figure 9).

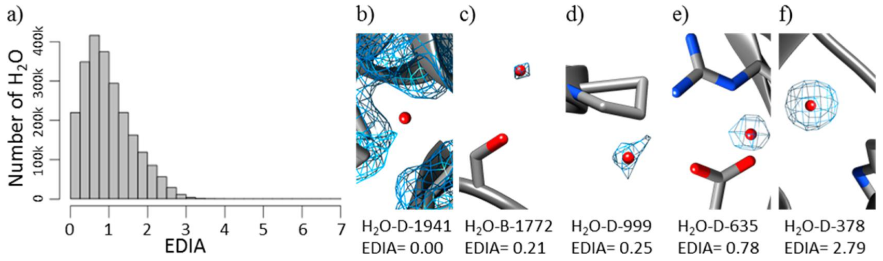


Figure 8. (a) EDIA histogram for all water molecules in the data set, mean = 0.981, median = 0.868, standard deviation = 0.625. (b−f) Examples of EDIA values and the corresponding visualization (3fpc, 1.4 Å). (b) Electron density map ($2f_o - f_c$ map, blue mesh) at 1σ for the whole region is shown. (c−f) $2f_o - f_c$ at 1σ is only shown for the water molecule itself (molecular graphics were created using UCSF Chimera[52]).

Table 5. Numbers of Water Molecules in Each Class[a] after Separation According to $EDIA_{Thrs}$

	well-resolved H_2O ($EDIA(H_2O) \geq EDIA_{Thrs}$)	insufficiently resolved H_2O ($EDIA(H_2O) < EDIA_{Thrs}$)
C	258106 (97.6%)	6478 (2.4%)
PLI	77673 (96.2%)	3044 (3.8%)
PPI	123,731 (96.7%)	4257 (3.3%)
S	1663019 (89.5%)	194273 (10.5%)

[a]Captured C, protein−ligand interface PLI, protein−protein interface PPI, surface S.

Table 6. Hydrogen Bonding Characteristics of Water Molecules[a]

	H_2O position	$Hbond_{total}$ ± stdev	$Hbond_{H2O}$ ± stdev
	all	2.15 ± 0.97	1.22 ± 0.93
H_2O EDIA ≥ $EDIA_{Thrs}$	all	2.12 ± 0.96	1.15 ± 0.92
	S	1.98 ± 0.94	1.19 ± 0.93
	C	2.70 ± 0.83	0.87 ± 0.79
	PLI	2.47 ± 0.92	1.03 ± 0.89
	PPI	2.60 ± 0.90	1.31 ± 0.94
H_2O EDIA < $EDIA_{Thrs}$	all	1.65 ± 0.89	1.14 ± 0.89
	S	1.62 ± 0.89	1.15 ± 0.89
	C	1.88 ± 0.96	0.76 ± 0.79
	PLI	2.03 ± 0.92	1.04 ± 0.88
	PPI	2.09 ± 0.95	1.24 ± 0.95

[a]Captured C, protein−ligand interface PLI, protein−protein interface PPI, surface S. $Hbond_{total}$ = all hydrogen bonds to protein. Ligand or other water molecules were considered. $Hbond_{H2O}$ = hydrogen bonds only to other water molecules were considered. stdev = one standard deviation.

Therefore, water molecules were further classified according to their position in the biological complex (S, C, PLI, PPI, see the Materials and Methods section), allowing a more detailed view on their hydrogen bonding characteristics (see Figure 9). As expected, in all four categories the mean number of hydrogen bonds ($Hbond_{total}$) decreases from resolved to unresolved water molecules (see Table 5). The most drastic decrease can be seen for unresolved captured water molecules, which form about one hydrogen bond less in comparison to resolved captured water molecules. These results show that modeled water molecules in positions without experimental proof are less integrated into the hydrogen bonding network.

The likelihood of "missing partners" is a bias of surface water molecules. Either further shells of water molecules are not resolved and remain unmodeled or they might interact with neighboring protein chains that are not in the asymmetric unit considered here. Analyzing the number of bulk water accessible hydrogen bonding functions, surface water molecules have a mean of 1.69 and 2.06 accessible hydrogen bonding functions for resolved and insufficiently resolved water molecules. These numbers might be slightly overestimated, given that we tested whether 75% of the volume of a water molecule would fit in the ideal direction of a hydrogen bonding function. In total this leads to 3.67 hydrogen bonds for both resolved as well as unresolved surface water molecules, close to the number of about 3.5 hydrogen bonds on average in bulk at 298 K.[56−58]

Hydrophobic Bubbles. Surprisingly, we found captured water molecules that are resolved by electron density ($mean_{EDIA}$ = 0.88 ± 0.56) but do not form any hydrogen bonds with protein, ligand atoms, or water molecules (Figures 9a and 10). These water molecules resemble so-called hydrophobic bubbles.[40] In total, only 1438 (0.54%) of all captured water molecules are hydrophobic bubbles sufficiently resolved by electron density (0.06% of the whole data set). They are highly constrained in their position inside the protein, which is probably the reason why they are resolved by electron density; but display a higher thermal motion than PPI, PLI, or other captured water molecules and about the same B-factor as surface water molecules (see Table 7). Since these hydrophobic bubbles must be highly energetically unfavorable, as they are spatially constrained and cannot compensate the enthalpic loss by hydrogen bond formation, they might present predetermined breaking points of protein structures. An alternative explanation for some of the hydrophobic bubbles would be that the electron density comes from a noble gas, which was used to solve the phase problem.[59−61] The electron density of a noble gas would be hardly differentiable from a water molecule especially if the position is only partially occupied.

Well-Integrated Unresolved Water Molecules. Another interesting result is given by water molecules that are not resolved (EDIA < $EDIA_{Thrs}$) but are very well integrated into a hydrogen bonding network forming three or four hydrogen bonds to protein or ligand (see Figure 11). In total 769 of those are found in our data set (referred to as $WIND_{PL}$). The number increases substantially to 34 217 if surrounding water molecules are considered as a hydrogen bonding partner (referred to as $WIND_{all}$). Simply accounting for the high number of formed hydrogen bonds, the water molecule would be expected to be resolved by electron density. However, there is little or no experimental proof for the water molecule to be in this place. One of the reasons for the lack of electron density might be their hydrogen bonding partner. Compared to resolved water molecules, $WIND_{all}$ water molecules build 13% to 40% more

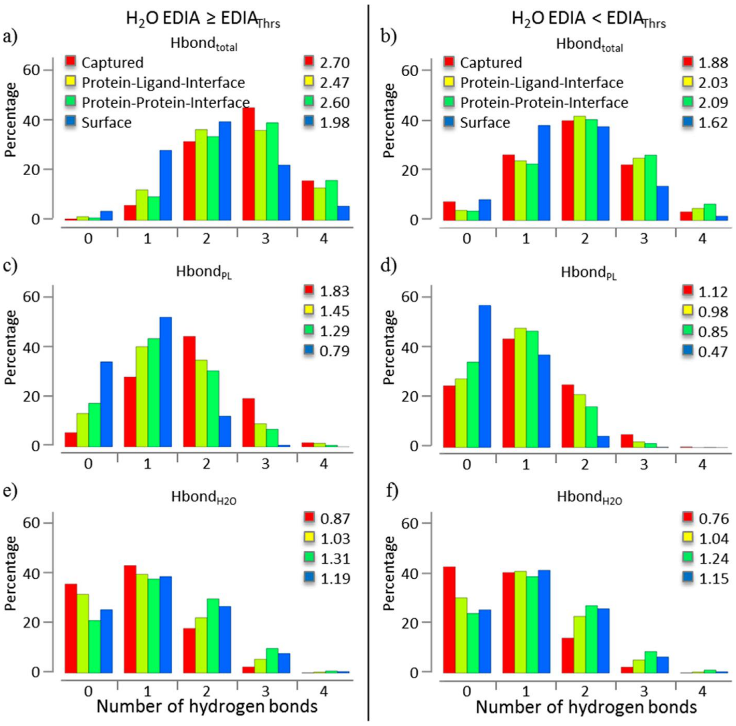


Figure 9. Histograms of hydrogen bonds formed by all water molecules from the data set. Data is separated for each class of water molecules, each class represents 100% of its water molecules. Hydrogen bonds of water molecules (a, c, e) with electron density (EDIA ≥ EDIA$_{Thrs}$) and (b, d, f) without electron density (EDIA < EDIA$_{Thrs}$). (a and b) All hydrogen bonds (to protein, ligand, and other water molecules) are counted. (c and d) Only hydrogen bonds to explicit partners, protein or ligand, are counted. (e and f) Only hydrogen bonds to other water molecules are counted. Numbers in the legend are the mean number of hydrogen bonds for each class.

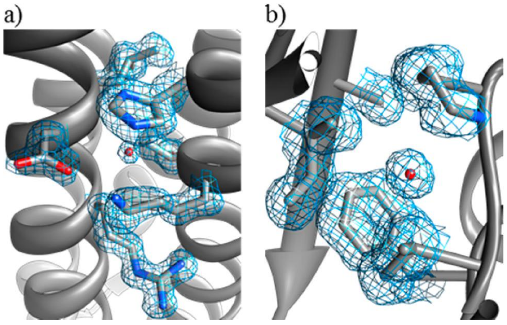


Figure 10. Examples for hydrophobic bubbles. (a) H_2O-I-1613, EDIA = 0.45 (PDB ID: 3ak8) and (b) H_2O-A-535, EDIA = 1.58 (PDB ID: 4h5i): blue = $2f_o - f_c$ map at 1σ (molecular graphics were created using UCSF Chimera[52]).

hydrogen bonds to other water molecules (see Table 8). These water molecules may not have a favored position, since they interact less often with amino acids, and their molecular motion might therefore be too high to be detected during X-ray crystallography (see Table 7). A closer look at the 769 WIND$_{PL}$ as well as WIND$_{all}$ water molecules and their surroundings often reveals highly flexible regions in which the water molecules are incorporated (see Table 7).

Hydrogen Bonding Partners and Proximity Preferences of Water Molecules. Classified water molecules were further analyzed concerning hydrogen bonding partner preferences and their favored surroundings.

First, we analyzed whether water molecules primarily interact with backbone or side chain functional groups. Therefore, we calculated the ratio of hydrogen bonding functions from backbone to side chain in our high-resolution data set. The distribution of backbone hydrogen bonding functions as opposed to side chain ones is shifted toward the backbone by

Table 7. Average B-Factor for Water Molecules with EDIA ≥ EDIA$_{Thrs}$ (total, S, C, PLI, PPI, HB), Well-Integrated Water Molecules with EDIA < EDIA$_{Thrs}$[a] and for All High-Resolution Proteins

	total	S	C	PLI	PPI	WIND$_{all}$	WIND$_{PL}$	HB	Protein
B-factor	27.42	29.35	18.35	22.71	23.40	44.03	45.60	28.97	15.66

[a] WIND$_{all}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; WIND$_{PL}$ ≥ three hydrogen bonds to protein or ligand. HB = hydrophobic bubbles.

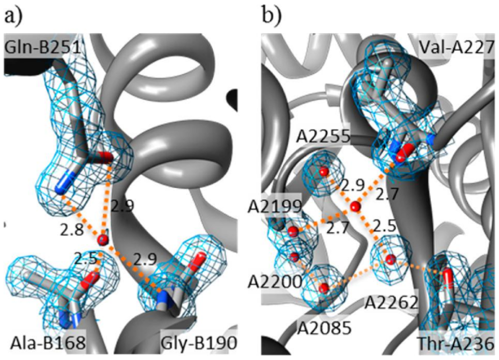


Figure 11. Examples for well-integrated, unresolved (EDIA < $EDIA_{Thrs}$) water molecules (a) H_2O-B-2135, EDIA = 0.16 (PDB ID: 4bgb) and (b) H_2O-A-2254, EDIA = 0.00 (PDB ID: 1of8); blue = $2f_o - f_c$ map at 1σ, orange dashed lines = hydrogen bonds with distances in angstroms (molecular graphics were created using UCSF Chimera[52]).

Table 8. Hydrogen Bonding Partners of Water Molecules with Sufficient Electron Density (C, PLI, PPI, S: EDIA ≥ $EDIA_{Thrs}$) Compared to Well-Integrated Unresolved Water Molecules[a]

	Hbond to H_2O	Hbond to PL
C	32.27%	67.73%
PLI	41.55%	58.45%
PPI	50.37%	49.63%
S	59.98%	40.02%
$WIND_{all}$	73.06%	26.94%
$WIND_{PL}$	6.93%	93.07%

[a] $WIND_{all}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; $WIND_{PL}$ ≥ three hydrogen bonds to protein or ligand; EDIA < $EDIA_{Thrs}$; Hbond = hydrogen bond; PL = hydrogen bond formed with either protein or ligand.

65%. Most of the hydrogen bonding functions of the protein backbone are satisfied due to their secondary structure patterns (α-helices and β-sheets). Therefore, it is not surprising, that PLI and PPI water molecules are more likely to interact with amino acid side chains (see Table 9). Especially PLI water molecules satisfy hydrogen bonding functions of side chains, whereas captured water molecules preferably interact with the protein backbone. The latter was also found in previous studies.[39,62] We found that PPI water molecules slightly favor side chain interaction (52%), however to a smaller extent than estimated by Ahmed et al. (78.5%)[40] but more than found by Rodier (45%)[10].

Second, atom-based preferences were dissected. It is noticeable that all water molecules are more likely to interact with oxygen atoms than the proportion of nitrogen to oxygen atoms of the high-resolution proteins would suggest. Similar to the backbone to side chain preferences, captured and PLI water molecules display the most extremes, with PLI water molecules interacting by 73% with oxygen atoms (See Table 9).

Third, since water molecules have two acceptor and donor functions each, we examined whether they do have any bias toward acceptor or donor interaction partners. Herein, the proportion is for all water molecules highly similar to the acceptor/donor distribution of the high-resolution proteins. Only PLI water molecules and well-integrated water molecules with insufficient electron density (referred to as $WIND_{all}$) show a greater bias toward acceptor interaction partners than water molecules from the other categories (see Table 9).

The last two results are directly connected to each other. Most of the time oxygen atoms are acceptors, wherein nitrogen atoms are more often present as donors. The proportion of oxygen acceptors in the high-resolution data set is nearly double the amount of nitrogen donors, 18 times the number of oxygen donors, and more than 100 times the amount of nitrogen acceptors. Therefore, it is most likely that water molecules interact with oxygen acceptors. Additionally, these results are in accordance with previous findings using quantum mechanical calculations.[63] Interactions from water molecules to either nitrogen or oxygen do not lead to significant differences in the estimated binding energy. However, water molecules interacting with acidic groups leads to an increase in binding affinity. This correlates well with the finding, that glutamate and aspartate are frequent interaction partners (see Table 10).

Given the different classes of water molecules we further investigated their functional group and amino acid preferences. Six different groups were formed, wherein each corresponds to one or more amino acid types (see Table 10). The preferences were then compared to the mean occurrence of amino acids in the high-resolution protein structures. Most noticeable is the high proportion of hydrogen bonds of water molecules to aspartate and glutamate residues (for C, PLI, and PPI). This probability is compared to the normal occurrence of aspartate and glutamate residues in proteins, which are the most abundant ones. Within protein−protein interfaces water molecules have a high probability to interact with arginine residues. Herein, the frequency is nearly as high as the normal occurrence of arginine residues within the protein. These findings are consistent with previous studies.[10,40] The results

Table 9. Hydrogen Bonding Partner Preferences of Water Molecules with Sufficient Density (total, C, PLI, PPI, S: EDIA ≥ $EDIA_{Thrs}$) Compared to Well-Integrated Unresolved Water Molecules[a] and the Respective Occurrence in the High-Resolution Proteins

	total	S	C	PLI	PPI	$WIND_{all}$	$WIND_{PL}$	protein
BB/SC	0.529	0.511	0.605	0.473	0.483	0.398	0.457	0.653
N/O	0.312	0.311	0.325	0.275	0.316	0.251	0.385	0.471
Don/Acc	0.325	0.322	0.341	0.293	0.329	0.254	0.396	0.363

[a] $WIND_{all}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; $WIND_{PL}$ ≥ three hydrogen bonds to protein or ligand; EDIA < $EDIA_{Thrs}$; BB/SC = ratio of backbone to side chain interactions; N/O = ratio of interactions to nitrogen or oxygen atoms of the protein, Don/Acc = ratio of hydrogen bonds to donor or acceptor functions of the protein. A ratio of 0.5 means equal distribution of the interaction partners. For the different categories of water molecules hydrogen bond partners are considered, wherein for the protein column available functions/atoms of the high-resolution data set are counted.

Table 10. Functional Group and Corresponding Amino Acid Preferences of Water Molecules with Sufficient Density (total, C, PLI, PPI, S: EDIA ≥ $\mathrm{EDIA_{Thrs}}$) Compared to Well-Integrated Unresolved Water Molecules[a] and the Respective Occurrence in the High-Resolution Proteins

interaction partner		total	S	C	PLI	PPI	$\mathrm{WIND_{all}}$	$\mathrm{WIND_{PL}}$	protein
functional group (amino acid)	amide (N, Q)	4	3	5	4	4	3	8	8
	imidazole/indole (H/W)	1	1	2	1	1	0	1	4
	amine (K)	2	2	1	2	2	2	8	6
	carboxyl (D, E)	7	7	8	8	8	6	13	12
	guanidine (R)	2	2	3	2	4	2	9	5
	hydroxyl (S, T, Y)	5	4	8	6	6	3	12	15
hydrophobic amino acids		–	–	–	–	–	–	–	50
ligand		1	0	0	16	0	0	0	–
H_2O		54	60	32	42	50	73	7	–
backbone		24	20	41	20	24	10	41	–

[a] $\mathrm{WIND_{all}}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; $\mathrm{WIND_{PL}}$ ≥ three hydrogen bonds to protein or ligand; EDIA < $\mathrm{EDIA_{Thrs}}$. For the water molecules hydrogen bond partners are considered (in percent), wherein for the protein the mean proportion of the respecting functional group is given (in percent). (−) No contribution.

Table 11. Proximity Preferences of Water Molecules with Sufficient Density (total, C, PLI, PPI, S: EDIA ≥ $\mathrm{EDIA_{Thrs}}$) Compared to Well-Integrated Unresolved Water Molecules[a]

	total	S	C	PLI	PPI	$\mathrm{WIND_{all}}$	$\mathrm{WIND_{PL}}$
hydrophobicity	0.617	0.613	0.639	0.607	0.625	0.594	0.609
hydrophobic surface	0.567	0.579	0.513	0.505	0.544	0.502	0.298
water cluster size	18	17	13	17	31	92	13

[a] $\mathrm{WIND_{all}}$ ≥ three hydrogen bonds to protein, ligand, or other water molecules; $\mathrm{WIND_{PL}}$ ≥ three hydrogen bonds to protein or ligand; EDIA < $\mathrm{EDIA_{Thrs}}$.

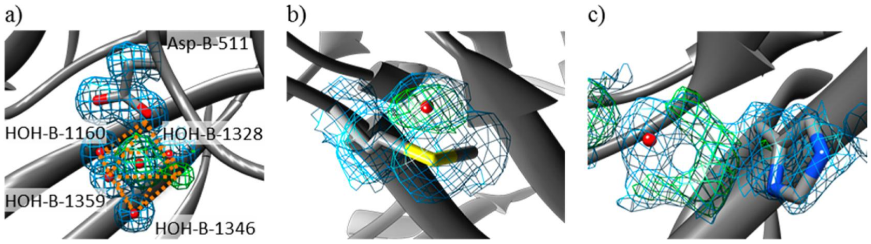


Figure 12. Modeling errors. (a) H_2O-B-1166 modeled into electron density that would better be suited for an ion (EDIA = 5.83, electron density difference map ($f_o - f_c$) indicates too little electrons modeled). Orange lines indicate possible ion coordination geometry (PDB ID: 1hyo). (b) H_2O-A-352 modeled into the electron density of methionine A-239 (distance 1.725 Å). The electron density difference map indicates missing electrons at the water position and too many electrons at the sulfur of methionine, EDIA = 4.38 (PDB ID: 2rdq). (c) H_2O-B-1328 modeled into alternative site chain conformation of Histidine B-985, EDIA = 1.73 (PDB ID: 1hyo). blue = $2f_o - f_c$ map at 1σ, green = $f_o - f_c$ map at 3σ (molecular graphics were created using UCSF Chimera[52]).

are in accordance with the above analysis, in which oxygen atoms as well as acceptors dominated the interaction partners of PLI water molecules. Only well-integrated water molecules with insufficient electron density (referred to as $\mathrm{WIND_{PL}}$) show a higher number of hydrogen bonds to carboxyl groups. Interestingly, the latter water molecules also have more interactions with lysine and arginine than their mean occurrence in the high-resolution protein structures (see Table 10 $\mathrm{WIND_{PL}}$). Another noticeable result is the small number of hydrogen bonds of surface water molecules with histidine or tryptophan, in accordance with previous studies that have shown that those amino acids are found less often on protein surfaces.[64,65]

Finally, water molecules were analyzed for proximity preferences as described in the Materials and Methods section. The hydrophobicity of their direct surrounding, their hydrophobic surface area, and the size of water clusters was examined (see Table 11). Noticeable is the relatively high proportion of hydrophobic atoms in the surrounding of captured water molecules, while the hydrophobic surface around them is comparably low. The latter characteristic allows captured water molecules to be well integrated into the protein complex. The difference between hydrophobicity and hydrophobic surface is even more remarkable, when looking at $\mathrm{WIND_{PL}}$ water molecules with 0.609 and 0.298 respectively. This shows, in accordance to the very high B-factor, that those areas are highly flexible (see Figure 11b).

Water clusters have on average 18 water molecules for all water molecules with an EDIA value greater than $\mathrm{EDIA_{Thrs}}$ (see Table 11). This number might appear quite large in comparison to previous studies.[66,67] However, previous analyses have focused on water clusters in cavities and not the whole protein. Additionally, the water clusters analyzed here have been detected based on a distance criterion only (see the Materials and Methods section). Interestingly, PPI water molecules show the biggest average number of water molecules within one water cluster among the four classes. Unsurprisingly, well-integrated, unresolved water molecules have a huge number of water molecules in one cluster if hydrogen bonds to water molecules are taken into account, relating again to their higher mobility.

Identification of Modeling Errors. Astonishingly, more than 1200 water molecules show very high EDIA values above median plus four standard deviations (EDIA > 3.3). Visual inspection of a random sample of 10% of those locations suggests modeling errors in over 75% (Figure 12a), with the electron density difference map showing too few electrons modeled in the position where the water molecule was placed. Most of those water molecules would better be substituted by ions, for which in at least 20% very good coordination geometries, such as octahedral or tetrahedral, can be found.

A further misinterpretation of the electron density was detected as water molecules may be built into the electron density of alternative amino acid configurations (Figure 12b and c). Identification of those modeling errors is more complex, as electron density supposed for other atoms is available. Automated identification of modeling errors or misinterpretation will be approached in future EDIA development.

CONCLUSION

Water molecules play an important role in many biological aspects, not only in mediating protein−ligand interactions, but also contributing fundamentally to binding affinity by dehydration and the hydrophobic effect. As those water molecules resolved by X-ray crystallography exist at local energy minima it would be advantageous to reliably predict those positions upon modeling molecular complexes.

In order to analyze the characteristics of water molecules a high-resolution subset consisting of 5485 structures from the PDB was compiled. Our evaluation has shown that high resolution itself is no guarantee for electron density support of each individual water molecule. Analyzing the electron density is unavoidable to differentiate between well resolved and unresolved water molecules. Therefore, a new measure based on electron density was developed, called EDIA. Advantages compared to already existing measurements, like B-factor, RSR, and RSCC are a direct comparison between modeled structure and electron density, as well as an intuitive interpretation of the value itself. Normalization by the standard deviation of the electron density map allows direct comparison of water molecules from different structures.

In order to detect misinterpretations of the electron density map and modeling errors the EDIA could be enhanced by taking the electron density difference map into account. In this way, further areas of too little or too many electrons in the modeled structure could be detected in an automated manner. Further water molecules misleadingly placed in electron density supposed for alternative amino acid or ligand conformations could be detected using a Difference of Gaussian filter.[68] Herein, two different sigma levels would be applied to the electron density map. Subtracting one image from the other preserves positions with drastic shifts, but discards all points that are at continuous areas, thus eliminating noise. As water molecules have a fairly circular, secluded distribution of electron density, it would become apparent if the underlying electron density, in which the water molecule is placed, is more stretched out and extensive as it is the case for amino acid side chains or ligands. Both aspects will be evaluated in our future development of EDIA.

The new measure EDIA can support the differentiation of water molecules that should be excluded from an analysis due to insufficient electron density support (EDIA < 0.24), from those that actually have implications for further modeling. Water molecules with unrealistically high EDIA values (EDIA > 3.3) need more attention due to a high probability of a wrong interpretation of the electron density. Even though very rarely, hydrophobic bubbles with good EDIA values were observed in this data set (1438 ≙ 0.52%) and showed that they may be of biological relevance. Otherwise, such highly unfavorable locations for water molecules would not be expected to exist. The observation from this evaluation provide further support for validating water molecules in crystal structures as well as implications for further characterization and modeling. In a subsequent analysis it would be highly interesting to investigate the underlying thermodynamics in order to understand why experimentally observed water molecules seem to be stable in their surroundings. This refers in particular to water molecules in a hydrophobic environment.

Many computational methods that aim to predict water molecule locations have been lately developed. However, little has been undertaken for the validation of water molecules in protein structures. As proven by the number of water molecules without electron density (208 052 ≙ 8.9% of the data set), a simple comparison with structurally modeled water molecules might not be sufficient. Herein, this well characterized high-resolution data set allows an extensive evaluation of water prediction methods, including the possibility to differentiate between water molecules well-resolved by electron density and those not supported by electron density.

In summary, the EDIA serves two purposes. First, properties and functions of meaningful modeled water molecules in crystal structures can be characterized and comprehended. Second, it can support the validation of computational methods for placing water molecules.

ASSOCIATED CONTENT

S Supporting Information

Table S1: Data set of all PDBids. Table S2: Table of captured, PLI, and PPI water molecules including EDIA and diverse descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The project is part of the Biokatalyse2021 cluster and is funded by the Federal Ministry of Education and Research (BMBF) under grant 031A1838. We would like to thank our cooperation partner BioSolveIT GmbH, especially Holger Claussen and Christian Lemmen for fruitful discussions. Special thanks to Robert Klein for helpful information about the electron density map format. We also thank Karen Schomburg for critical reading and revising the manuscript.

ABBREVIATIONS

EDIA, electron density for individual atoms; $2f_o - f_c$, composite electron density map; $f_o - f_c$, electron density difference map; MSA, molecular surface area; PLI, protein−ligand interface; PPI, protein−protein interface; RSCC, real-space correlation coefficient; RSR, real-space R-factor

REFERENCES

(1) Timasheff, S. N. The Control of Protein Stability and Association by Weak Interactions with Water: How Do Solvents Affect These Processes? *Annu. Rev. Biophys. Biomol. Struct.* **1993**, *22*, 67−97.

(2) Zhang, L.; Yang, Y.; Kao, Y.-T.; Wang, L.; Zhong, D. Protein Hydration Dynamics and Molecular Mechanism of Coupled Water-Protein Fluctuations. *J. Am. Chem. Soc.* **2009**, *131*, 10677−10691.

(3) Levy, Y.; Onuchic, J. N. Water Mediation in Protein Folding and Molecular Recognition. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 389−415.

(4) Mattos, C. Protein−water Interactions in a Dynamic World. *Trends Biochem. Sci.* **2002**, *27*, 203−208.

(5) Ahmad, S.; Kamal, M. Z.; Sankaranarayanan, R.; Rao, N. M. Thermostable Bacillus Subtilis Lipases: In Vitro Evolution and Structural Insight. *J. Mol. Biol.* **2008**, *381*, 324−340.

(6) Kawasaki, K.; Kondo, H.; Suzuki, M.; Ohgiya, S.; Tsuda, S. Alternate Conformations Observed in Catalytic Serine of Bacillus Subtilis Lipase Determined at 1.3 Å Resolution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 1168−1174.

(7) Rühlmann, A.; Kukla, D.; Schwager, P.; Bartels, K.; Huber, R. Structure of the Complex Formed by Bovine Trypsin and Bovine Pancreatic Trypsin Inhibitor. *J. Mol. Biol.* **1973**, *77*, 417−436.

(8) Phillips, R. S. How Does Active Site Water Affect Enzymatic Stereorecognition? *J. Mol. Catal. B Enzym.* **2002**, *19−20*, 103−107.

(9) Langhorst, U.; Backmann, J.; Loris, R.; Steyaert, J. Analysis of a Water Mediated Protein−Protein Interactions within RNase T1. *Biochemistry* **2000**, *39*, 6586−6593.

(10) Rodier, F.; Bahadur, R. P.; Chakrabarti, P.; Janin, J. Hydration of Protein-Protein Interfaces. *Proteins* **2005**, *60*, 36−45.

(11) Janin, J. Wet and Dry Interfaces: The Role of Solvent in Protein−protein and protein−DNA Recognition. *Structure* **1999**, *7*, R277−R279.

(12) Ahmad, M.; Gu, W.; Geyer, T.; Helms, V. Adhesive Water Networks Facilitate Binding of Protein Interfaces. *Nat. Commun.* **2011**, *2*, 261.

(13) Chothia, C.; Janin, J. Principles of Protein−protein Recognition. *Nature* **1975**, *256*, 705−708.

(14) Young, L.; Jernigan, R. L.; Covell, D. G. A Role for Surface Hydrophobicity in Protein-Protein Recognition. *Protein Sci.* **1994**, *3*, 717−729.

(15) Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Studies of Protein-Protein Interfaces: A Statistical Analysis of the Hydrophobic Effect. *Protein Sci.* **1997**, *6*, 53−64.

(16) Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, *437*, 640−647.

(17) Shimokhina, N.; Bronowska, A.; Homans, S. W. Contribution of Ligand Desolvation to Binding Thermodynamics in a Ligand-Protein Interaction. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 6374−6376.

(18) Snyder, P. W.; Mecinovic, 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. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 17889−17894.

(19) 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. Chem., Int. Ed. Engl.* **2013**, *52*, 1822−1828.

(20) 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. *Chem. Biol.* **1996**, *3*, 973−980.

(21) Biela, A.; Khayat, M.; Tan, H.; Kong, J.; Heine, A.; Hangauer, D.; Klebe, G. Impact of Ligand and Protein Desolvation on Ligand Binding to the S1 Pocket of Thrombin. *J. Mol. Biol.* **2012**, *418*, 350−366.

(22) Chen, J. M.; Xu, S. L.; Wawrzak, Z.; Basarab, G. S.; Jordan, D. B. Structure-Based Design of Potent Inhibitors of Scytalone Dehydratase: Displacement of a Water Molecule from the Active Site. *Biochemistry* **1998**, *37*, 17735−17744.

(23) Wissner, A.; Berger, D. M.; Boschelli, D. H.; Floyd, M. B.; Greenberger, L. M.; Gruber, B. C.; Johnson, B. D.; Mamuya, N.; Nilakantan, R.; Reich, M. F.; Shen, R.; Tsou, H.-R.; Upeslacis, E.; Wang, Y. F.; Wu, B.; Ye, F.; Zhang, N. 4-Anilino-6,7-Dialkoxyquinoline-3-Carbonitrile Inhibitors of Epidermal Growth Factor Receptor Kinase and Their Bioisosteric Relationship to the 4-Anilino-6,7-Dialkoxyquinazoline Inhibitors. *J. Med. Chem.* **2000**, *43*, 3244−3256.

(24) Seo, J.; Igarashi, J.; Li, H.; Martasek, P.; Roman, L. J.; Poulos, T. L.; Silverman, R. B. Structure-Based Design and Synthesis of N(omega)-Nitro-L-Arginine-Containing Peptidomimetics as Selective Inhibitors of Neuronal Nitric Oxide Synthase. Displacement of the Heme Structural Water. *J. Med. Chem.* **2007**, *50*, 2089−2099.

(25) Zhang, L.; Hermans, J. Hydrophilicity of Cavities in Proteins. *Proteins* **1996**, *24*, 433−438.

(26) 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*, 172−182.

(27) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **2007**, *129*, 2577−2587.

(28) 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*, 1063−1067.

(29) Adams, D. J. Grand Canonical Ensemble Monte Carlo for a Lennard-Jones Fluid. *Mol. Phys.* **1975**, *29*, 307−311.

(30) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(31) Grant, J. A.; Pickup, B. T.; Nicholls, A. A Smooth Permittivity Function for Poisson-Boltzmann Solvation Methods. *J. Comput. Chem.* **2001**, *22*, 608−640.

(32) 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*, 2817−2831.

(33) 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.

(34) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Prediction of the Water Content in Protein Binding Sites. *J. Phys. Chem. B* **2009**, *113*, 13337−13346.

(35) Nguyen, C. N.; Young, T. K.; Gilson, M. K. Grid Inhomogeneous Solvation Theory: Hydration Structure and Thermodynamics of the Miniature Receptor cucurbit[7]uril. *J. Chem. Phys.* **2012**, *137*, 044101.

(36) 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*, e32036.

(37) Poornima, C. S.; Dean, P. M. Hydration in Drug Design. 1. Multiple Hydrogen-Bonding Features of Water Molecules in Mediating Protein-Ligand Interactions. *J. Comput. Aided. Mol. Des.* **1995**, *9*, 500−512.

(38) 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*, 668−675.

(39) Park, S.; Saven, J. G. Statistical and Molecular Dynamics Studies of Buried Waters in Globular Proteins. *Proteins* **2005**, *60*, 450−463.

(40) Ahmed, M. H.; Spyrakis, F.; Cozzini, P.; Tripathi, P. K.; Mozzarelli, A.; Scarsdale, J. N.; Safo, M. A.; Kellogg, G. E. Bound

Water at Protein-Protein Interfaces: Partners, Roles and Hydrophobic Bubbles as a Conserved Motif. *PLoS One* **2011**, *6*, e24712.

(41) Dunitz, J. D. The Entropic Cost of Bound Water in Crystals and Biomolecules. *Science* **1994**, *264*, 670.

(42) Huggins, D. J. Benchmarking the Thermodynamic Analysis of Water Molecules around a Model Beta Sheet. *J. Comput. Chem.* **2012**, *33*, 1383−1392.

(43) Huggins, D. J. Application of Inhomogeneous Fluid Solvation Theory to Model the Distribution and Thermodynamics of Water Molecules around Biomolecules. *Phys. Chem. Chem. Phys.* **2012**, *14*, 15106−15117.

(44) 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*, 871−883.

(45) Kinoshita, M. Importance of Translational Entropy of Water in Biological Self-Assembly Processes like Protein Folding. *Int. J. Mol. Sci.* **2009**, *10*, 1064−1080.

(46) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in These Models. *Acta Crystallogr. Sect. A Found. Crystallogr.* **1991**, *47*, 110−119.

(47) Hawkins, P. C. D.; Kelley, B. P.; Warren, G. L. The Application of Statistical Methods to Cognate Docking: A Path Forward? *J. Chem. Inf. Model.* **2014**, *54*, 1339−1355.

(48) Levitt, M.; Park, B. H. Water: Now You See It, Now You Don't. *Structure* **1993**, *1*, 223−226.

(49) 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 Res.* **2000**, *28*, 235−242.

(50) Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wählby, A.; Jones, T. A. The Uppsala Electron-Density Server. *Acta Crystallogr. D. Biol. Crystallogr.* **2004**, *60*, 2240−2249.

(51) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminform.* **2014**, *6*, 12.

(52) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera–a Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

(53) Ko, T.-P.; Robinson, H.; Gao, Y.-G.; Cheng, C.-H. C.; DeVries, A. L.; Wang, A. H.-J. The Refined Crystal Structure of an Eel Pout Type III Antifreeze Protein RD1 at 0.62-A Resolution Reveals Structural Microheterogeneity of Protein and Solvation. *Biophys. J.* **2003**, *84*, 1228−1237.

(54) Karplus, P. A.; Faerman, C. Ordered Water in Macromolecular Structure. *Curr. Opin. Struct. Biol.* **1994**, *4*, 770−776.

(55) Carugo, O.; Bordo, D. How Many Water Molecules Can Be Detected by Protein Crystallography? *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1999**, *55*, 479−483.

(56) Hoffmann, M. M.; Conradi, M. S. Are There Hydrogen Bonds in Supercritical Water? *J. Am. Chem. Soc.* **1997**, *119*, 3811−3817.

(57) Soper, A. K.; Bruni, F.; Ricci, M. A. Site−site Pair Correlation Functions of Water from 25 to 400 °C: Revised Analysis of New and Old Diffraction Data. *J. Chem. Phys.* **1997**, *106*, 247.

(58) Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odelius, M.; Ogasawara, H.; Näslund, L. A.; Hirsch, T. K.; Ojamäe, L.; Glatzel, P.; Pettersson, L. G. M.; Nilsson, A. The Structure of the First Coordination Shell in Liquid Water. *Science* **2004**, *304*, 995−999.

(59) Tilton, R. F.; Kuntz, I. D.; Petsko, G. A. Cavities in Proteins: Structure of a Metmyoglobin Xenon Complex Solved to 1.9.ANG. *Biochemistry* **1984**, *23*, 2849−2857.

(60) Prangé, T.; Schiltz, M.; Pernot, L.; Colloc'h, N.; Longhi, S.; Bourguet, W.; Fourme, R. Exploring Hydrophobic Sites in Proteins with Xenon or Krypton. *Proteins* **1998**, *30*, 61−73.

(61) Schiltz, M.; Fourme, R.; Prangé, T. Use of Noble Gases Xenon and Krypton as Heavy Atoms in Protein Structure Determination. *Methods Enzymol.* **2003**, *374*, 83−119.

(62) Williams, M. A.; Goodfellow, J. M.; Thornton, J. M. Buried Waters and Internal Cavities in Monomeric Proteins. *Protein Sci.* **1994**, *3*, 1224−1235.

(63) Rezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7*, 2427−2438.

(64) Miller, S.; Janin, J.; Lesk, A. M.; Chothia, C. Interior and Surface of Monomeric Proteins. *J. Mol. Biol.* **1987**, *196*, 641−656.

(65) Fukuchi, S.; Nishikawa, K. Protein Surface Amino Acid Compositions Distinctively Differ between Thermophilic and Mesophilic Bacteria. *J. Mol. Biol.* **2001**, *309*, 835−843.

(66) Yin, H.; Hummer, G.; Rasaiah, J. C. Metastable Water Clusters in the Nonpolar Cavities of the Thermostable Protein Tetrabrachion. *J. Am. Chem. Soc.* **2007**, *129*, 7369−7377.

(67) Vaitheeswaran, S.; Yin, H.; Rasaiah, J. C.; Hummer, G. Water Clusters in Nonpolar Cavities. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 17002−17005.

(68) Marr, D.; Hildreth, E. Theory of Edge Detection. *Proc. R. Soc. London B. Biol. Sci.* **1980**, *207*, 187−217.

Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures.

[D3] Meyder, A.; **Nittinger, E.**; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. J. Chem. Inf. Model. 2017, 57 (10): 2437-2447.

http://pubs.acs.org/articlesonrequest/AOR-uPFmKQEyZHYbXYyxP4WE

Journal of
CHEMICAL INFORMATION
AND MODELING

Article

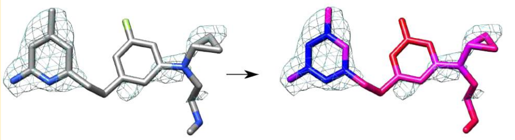
Cite This: J. Chem. Inf. Model. 2017, 57, 2437-2447 pubs.acs.org/jcim

Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures

Agnes Meyder,[†] Eva Nittinger,[†] Gudrun Lange,[‡] Robert Klein,[‡] and Matthias Rarey*,[†]

[†]ZBH—Center for Bioinformatics, Universität Hamburg, Hamburg 20146, Germany
[‡]Bayer AG, Frankfurt 65929, Germany

S Supporting Information


Ligand S90 of 4ugl with EDIA_m= 0.24 shown in Element and EDIA coloring.

ABSTRACT: Macromolecular structures resolved by X-ray crystallography are essential for life science research. While some methods exist to automatically quantify the quality of the electron density fit, none of them is without flaws. Especially the question of how well individual parts like atoms, small fragments, or molecules are supported by electron density is difficult to quantify. While taking experimental uncertainties correctly into account, they do not offer an answer on how reliable an individual atom position is. A rapid quantification of this atomic position reliability would be highly valuable in structure-based molecular design. To overcome this limitation, we introduce the electron density score EDIA for individual atoms and molecular fragments. EDIA assesses rapidly, automatically, and intuitively the fit of individual as well as multiple atoms (EDIA_m) into electron density accompanied by an integrated error analysis. The computation is based on the standard $2fo - fc$ electron density map in combination with the model of the molecular structure. For evaluating partial structures, EDIA_m shows significant advantages compared to the real-space R correlation coefficient (RSCC) and the real-space difference density Z score (RSZD) from the molecular modeler's point of view. Thus, EDIA abolishes the time-consuming step of visually inspecting the electron density during structure selection and curation. It supports daily modeling tasks of medicinal and computational chemists and enables a fully automated assembly of large-scale, high-quality structure data sets. Furthermore, EDIA scores can be applied for model validation and method development in computer-aided molecular design. In contrast to measuring the deviation from the structure model by root-mean-squared deviation, EDIA scores allow comparison to the underlying experimental data taking its uncertainty into account.

INTRODUCTION

Protein crystal structures are essential in gaining insights into biochemical processes. Depending on the quality of experimental data, the positions of individual atoms are regularly determined. In an iterative refinement process with assistance from the proposed molecular model, the final model is derived from the measured reflection intensities.[1] In many cases, the resulting model is the starting point of structure exploitation neglecting the gap toward the experimentally determined electron density map. In the past, measurements to assess the quality of protein structure models originating from X-ray crystallography on the global as well as on the local scale have been intensely discussed.[2−4] While global measures such as the diffraction precision index (DPI)[5] and the R factor[6] exist, analyzing and measuring the experimental support for substructures or regions in the structure such as binding pockets and ligands is still an active field of research.[3,7] Atoms in the protein structure models are usually annotated with occupancy and B factor. Both are calculated based on the initial structure factors as a step in the overall protein refinement procedure. The occupancy is set to less than one if the atom has an alternate location. The B factor mirrors local motion at the respective location. Since the B factor is a value optimized by the refinement program including user and program dependent constraints,[8−10] it can only be calculated for atom models, present at the refinement stage of the structure and is tightly coupled to the aforementioned constraints.

The analysis of electron density maps offers an alternative path for structure validation. The $2fo - fc$ and the $fo - fc$ electron density (difference) maps are often used for visual inspection. The fo map contains the experimentally observed electron density, and the fc map the calculated density derived from the structure model. The comparison of the fo to the fc map is used in different flavors. The analysis of the (squared) sum of errors $fo - fc$ is the base for the two well-known real-space refinement methods[11,12] implemented in, e.g. Coot,[13] Moe 2015.10,[14] and ARP/wARP.[15] As advancement, the real-space R factor (RSR) was proposed by Jones et al.[16] in 1991. The RSR compares the observed against the expected electron density around a specific atom or a group of atoms with the sum as the normalization factor (eq 1). It ranges from 0 to 1 in which 0 stands for a good density overlap. The RSR publication does not give definitions of all necessary parameters like the fitted scaling factor and the size of the electron density spheres required to calculate the expected density.

Received: June 27, 2017
Published: October 5, 2017

DOI: 10.1021/acs.jcim.7b00391

$$\mathrm{RSR(area)} = \frac{\sum |\rho_{\mathrm{obs}} - \rho_{\mathrm{calc}}|}{\sum |\rho_{\mathrm{obs}} + \rho_{\mathrm{calc}}|} \quad (1)$$

Consequently, reimplementations of the RSR differ in the chosen radii as well as the method to compute the expected electron density. In general, the RSR is resolution dependent due to the unspecified minimum grid spacing of the electron density map. Additionally, the fitting of the scaling factor can introduce an error.[17] RSR in its normalized form (RSR-Z) is used in the PDB structure validation pipeline. Unfortunately, ligands can not be checked thoroughly due to missing statistical data necessary for RSR-Z calculation.[3] RSR's correlation coefficient (RSCC)[18] was developed subsequently and is the most commonly used comparison metric so far. As a correlation coefficient, it ranges from −1 to 1. The RSCC avoids the need for an electron density scaling factor. Still, the other problematic issues of the RSR prevail. Thus, the RSR and RSCC are problematic when comparing structures with different resolutions as well as the same structure scored via different programs. In the context of structure comparison, a variation of the RSR, called the RSR$_n$,[19] was developed. The RSR$_n$ scales the RSR of a structure model against the RSR of the crystallized ligand. By avoiding the resolution dependency, the RSR$_n$ allows better comparison between structures of different resolutions if the crystallized ligand is suitable for normalization. Nevertheless, all three methods do not check for clashing atoms, do not report the intensity of the local electron density and do not consider unaccounted density beyond the electron density radius of the atom. Thus, they can not detect a nonfitting electron density shape.[3] A recent development, the real-space difference density Z score[17] (RSZD) reports significant difference density peaks in the electron density sphere of an atom. It employs consecutive significance tests on the difference map. The RSZD can be split into the RSZD+ and RSZD− to ease the interpretation of reported outliers. RSZD− (below -3σ) marks areas with modeled atoms but no supporting density while RSZD+ (above 3σ) marks areas with unmodeled electron density. As RSZD measures how well the model fits the experimental data, the accompanying RSZO measures the precision of the model. It calculates the signal-to-noise ratio marking the regions well-defined by electron density with values larger than 1σ. The effects of *fc* on the RSCC, RSZD, and RSZO can be startling due to the methods used and values such as annotated B factors and occupancies integrated in the calculation of *fc*. In addition, an electron density fit calculation in high throughput mode as applicable in molecular modeling should aim to avoid recalculation of, i.e., *fc* for a $fo - fc$ map per ligand conformer as much as possible. When curating a validation data set such as the Astex,[20] Iridium,[21] CSAR 2012,[22] and PDBbind core set 2013,[23] visual inspection, in combination with the RSCC or a density correlation coefficient based on the $2fo - fc$ and *fc* map[24] has been used. But since available high quality data has rapidly increased, a manual inspection is not feasible anymore and an automated evaluation procedure is needed.

In this paper we introduce a novel approach of the electron density support for individual atoms (EDIA) and its combination for multiple atoms (EDIA$_\mathrm{m}$) in crystal structures. A preliminary version of EDIA was developed by Nittinger et al.[25] to score the quality of water molecules. While the initial version was limited to oxygen atoms of water molecules, we now present a generic concept allowing the analysis of any given atom or structure for its fit into the electron density while avoiding the aforementioned pitfalls. Both $2fo - fc$ and *fo* maps can be used in its calculation. In the following, the computation of EDIA is outlined and tested toward numerical stability. We then apply EDIA$_\mathrm{m}$ with the help of $2fo - fc$ maps from the PDBe[26] on ligands and residues in a high resolution PDB subset. Then we analyze EDIA on the Astex Diverse Set. EDIA and EDIA$_\mathrm{m}$ are subsequently compared to RSCC, RSZD, RSZO, B factor, and RMSD. The comparison to RMSD is of particular interest for evaluating sets of molecular models such as those resulting from docking calculations.

■ METHOD

EDIA is a method estimating the electron density support for an individual atom in a crystal structure. An EDIA value ranges from 0−1.2, the upper bound results from the truncation of the density score as explained later on. EDIA assumes a spherical shaped electron density for heavy atoms. As hydrogens rarely have electron density spheres themselves, they are per default excluded from the calculation. For the EDIA description, we will use the following nomenclature and abbreviations: $s(a)$ represents the electron density radius sphere around atom a. The sphere with two times the electron density radius is called *sphere of interest* of atom a. The surrounding of atom a defined as the difference between the sphere of interest and $s(a)$ for atom a is called $d(a)$. For a well-resolved atom $s(a)$ is expected to be filled with electron density while $d(a)$ should not contain unaccounted electron density. The calculation is divided into three phases: **(1)** the oversampling of the raw electron density grid, **(2)** the EDIA value calculation itself, and **(3)** the combination of EDIA values into overall values for whole molecules or molecular fragments. The oversampling procedure increases the density of grid points to guarantee at lest 27 grid points in $s(a)$ for every atom a. Details of this step can be found in SI 2.1.2

The EDIA calculation itself can be summarized as follows: After assigning an electron density radius to each atom a, three values are calculated for each grid point p in the sphere of interest of a: a distance-dependent weighting factor $w(p, a)$, an ownership value $o(p, a)$ reflecting the interference of neighboring atoms, and the density score $z(p)$ truncated at 1.2σ in which σ denotes the root-mean-square (RMS) value of the electron density.[27] Since the mean electron density is approximately zero, the RMS is roughly equal to the standard deviation of the map. The truncation balances the often very high density values in $s(a)$ against those in $d(a)$ so that density found in $d(a)$ has full penalty strength. The EDIA value for atom a is then the weighted mean over the product $w(p, a)o(p, a)z(p)$ over all assigned grid points (see eq 2). The calculation of $w()$ and $o()$ will be summarized below; additional information can be found in SI 2.

$$\mathrm{EDIA}(a) = \frac{\sum^{p \in M_{2fo-fc}} w(p, a)o(p, a)z(p)}{\sum_{p \in M_{2fo-fc} | w(p,a)>0} w(p, a)} \quad (2)$$

$\overline{pa} = \|p - a\|_2$ (distance)

$w(p, a)$: weight function depending on the distance $\overline{pa}$ (see below)

$o(p, a)$: ownership of p from a (see below)

$$z(p) = \begin{cases} 0 & \text{if } \dfrac{\rho(p) - \mu}{\sigma} < 0.0 \\ \dfrac{\rho(p) - \mu}{\sigma} & \text{if } 0 \leq \dfrac{\rho(p) - \mu}{\sigma} \leq \zeta \\ \zeta & \text{if } \dfrac{\rho(p) - \mu}{\sigma} > \zeta \end{cases}$$

$\zeta = 1.2$

$\rho(p)$: density at p

μ: mean of the $2fo - fc$ map

σ: root mean square of the $2fo - fc$ map

Atom Radius Determination. Tickle[17] approximates the electron density radius r for an atom by calculating the radius integral over the atom density. In his method, r depends on the B factor, element, charge, and structure resolution. Our aim is to mark overly flexible atoms as problematic to flag uncertain atomic positions. Combined with the problematic constrained optimization of B factors,[8−10] we decide to avoid the usage of the annotated atomic B factors. Instead, we analyze the B factors of all PDB complexes with ligands per resolution interval from 0.5 to 3.0 Å with a step size of 0.5 Å. The mean B factor per resolution interval is then used in the calculation of r. Six resulting electron density radii per element, charge, and resolutions of 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 Å are tabularized for fast lookup. The radius r for an atom is then linearly interpolated based on the resolution of the overall structure and the atom's charge. We assume a spherical shape for the electron density when computing the EDIA. Hence, we suggest to only compute EDIA values for structures with a resolution of at least 2.0 Å. The atomic scattering factors in an experiment with a resolution worse than 2.0 Å can not be approximated by the Gaussian functions anymore.[28] Nevertheless, EDIA is parametrized up to a resolution of 3.0 Å to allow the evaluation of older validation data sets such as the Iridium set.[21] The resulting atom radii are given in the SI 2.1.1.

Point Weighting. With the help of a weighting curve (Figure 1), EDIA can discriminate between necessary and superfluous electron density in the vicinity of an atom. If the grid point p is in $d(a)$, the density is regarded as superfluous and weighted negatively for a. If p is outside of the sphere of interest of a, the weight is set to 0. The weighting curve consists of three connected parabolas P1, P2, and P3 to be numerically stable and efficient to compute. The full parametrization of the weighting curve can be found in SI 2.1.3.

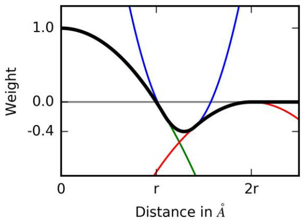


Figure 1. Weighting curve $w(p, a)$. A weight is assigned to a grid point p in the sphere of interest of atom a based on its distance to the atom center of a. The weighting curve, depending on the electron density radius r of a, shown in black, is composed of three parabolas P1 (green), P2 (blue), and P3 (red). The weight turns to zero beyond the sphere of interest.

Grid Point Ownership. In order to model the molecular structure, we define an ownership function $o(p, a)$, assigning grid points to atoms. Due to $o(p, a)$, EDIA is able to handle covalently bound atoms, correctly identify atomic clashes and superfluous density in the vicinity of an atom. The calculation is based on three sets defined for a grid point p: Let $S(p)$ be all atoms containing p in their electron density sphere, $D(p)$ all atoms containing p in their sphere of interest excluding $S(p)$. Furthermore, we define $I(p, a) \subseteq S(p)$ omitting all atoms covalently bound to a. Note that $I(p, a)$ contains the atom a itself. For calculating $o(p, a)$ we distinguish four cases as shown in Figure 2. Let us first assume $a \in S(p)$. In case a is the only atom in $I(p, a)$, the density fully belongs to a and $o(p, a)$ is set to 1.0. Otherwise, the ownership is shared between the atoms in $I(p, a)$. Note that due to the construction of $I(p, a)$, if $S(p)$ contains covalently bound atoms, p is fully assigned to all of them. Let us now assume $a \in D(p)$. In case $S(p)$ is not empty, the density belongs to a different atom and $o(p, a)$ is set to 0. Otherwise, the ownership is shared between the atoms in $D(p)$. Finally, in case a is neither in $S(p)$ nor in $D(p)$, the ownership is set to 0. For sharing the ownership within a set of atoms X, weights are calculated incorporating the distance of p to the atoms in X (see eq 3):

$$o(p, a) = \begin{cases} 1 & \text{if } |X| = 1 \\ 1 - \dfrac{\overline{pa}}{\sum_{b \in X} \overline{pb}} & \text{otherwise} \end{cases} \tag{3}$$

The resulting ownership decision tree is shown in Figure 2.

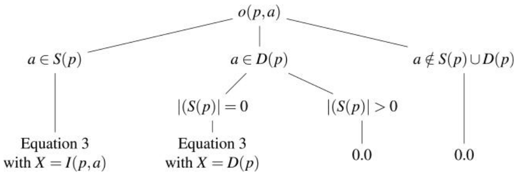

$o(p,a)$
$a \in S(p)$
$a \in D(p)$
$a \notin S(p) \cup D(p)$
$|(S(p)| = 0$
$|(S(p)| > 0$
Equation 3 with $X = I(p,a)$
Equation 3 with $X = D(p)$
0.0
0.0

Figure 2. Ownership decision tree for grid point p in the sphere of interest of atom a. If p lies outside of the sphere of interest of a, $o(p, a)$ is 0.0. $S(p)$: set of atoms containing p in $s(b)$. $D(p)$: set of atoms containing p in $d(b)$. $I(p, a) \subseteq S(p)$: set of atoms not covalently bound to a.

Identification of Error Types. EDIA can be used to detect problematic parts in a structure by quantifying the fit of the structure to the given electron density. Furthermore, it can be decomposed to automatically identify the two most frequently occurring error types: electron density sphere clashes and missing as well as unaccounted density. Also, the combination of both as possibly shifted electron density can be identified. An electron density sphere clash is detected when another atom b shares more than 10% of all grid points with a in its inner electron density sphere $s(a)$ (eq 4). An example can be found in Figure 3. In a distance of 1.66 Å, magnesium is positioned next to Oxygen 2 B in 2′-Deoxyguanosine-5′-Diphosphate (DGI) in the Nucleoside Diphosphate Kinase (3b6b[29]). Because the atoms have in this case an electron density sphere of 1.26 (Mg^{2+}) and 1.38 Å (O), the electron density spheres overlap profoundly.

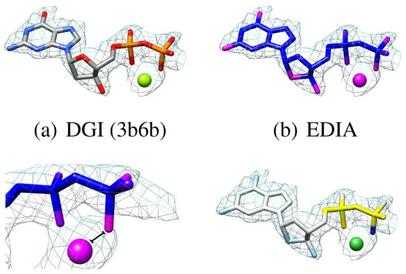


Figure 3. 2′-Deoxyguanosine-5′-diphosphate (DGI) in 3b6b as an example for all three errors detected by EDIA. Magnesium 138 D and Oxygen 2 B are recognized as clashing with a distance of 1.66 Å and electron density radii of r_{Mg}:1.26 Å and r_O:1.35 Å. All atoms are shown in element, EDIA (see Figure 5), and EDIA error coloring. EDIA error analysis color code: white no error; yellow density in $d(a)$; light blue not enough density in $s(a)$; blue clash, not enough density in $s(a)$; lime green clash, too much density in $d(a)$. The $2fo - fc$ map is shown at a contour level of 1σ.

$$\text{clash}(a, b) = \frac{2|\{p \in s(a) \cap s(b)\}|}{|\{p \in s(a)\}| + |\{p \in s(b)\}|} > 0.1 \tag{4}$$

Unaccounted density is detected in the vicinity of an atom a when the negatively weighted part of the EDIA is above 0.2 (eq 5, Figure 3d: atoms in yellow and lime green).

$$\text{EDIA}(a)_- = \frac{\sum^{p \in M_{2fo-fc}|w(p,a)<0} w(p, a)o(p, a)z(p)}{\sum_{p \in M_{2fo-fc}|w(p,a)<0} w(p, a)} > 0.2 \tag{5}$$

When the positively weighted part of the EDIA falls below 0.8, not enough electron density is detected (see eq 6, Figure 3d: atoms in light blue and blue). The score is equivalent with a sphere of radius $0.81r$ filled with electron density of 1σ around atom a at the resolution of 1 Å.

$$\text{EDIA}(a)_+ = \frac{\sum^{p \in M_{2fo-fc}|w(p,a)>0} w(p, a)o(p, a)z(p)}{\sum_{p \in M_{2fo-fc}|w(p,a)>0} w(p, a)} < 0.8 \tag{6}$$

If both unaccounted and missing density are detected, it is labeled as possibly shifted electron density.

EDIA$_m$. We suggest eq 7 based on the power mean to compute the EDIA$_m$ value for a set of atoms U in regard to electron density fit.

$$\text{EDIA}_m(U) = \left(\frac{1}{|U|} \sum_{a \in U} (\text{EDIA}(a) + 0.1)^{-2} \right)^{-1/2} - 0.1 \tag{7}$$

Herein, the temporary correction of +0.1 safeguards the EDIA$_m$ value against overweighting a single EDIA value close to zero. In principle many functional forms could be used to accumulate EDIA values over atom collections. The power mean behaves like a soft minimum function which we feel appropriate to detect unsupported structural features in small molecules and fragments like amino acids. The tendency to reflect the minimum of a given set of scores strongly increases with an increasing negative deviation from the score set mean. The EDIA$_m$ results in a value below 0.8 if about one EDIA value drops below 0.3 or three EDIA drop values below 0.35 (see Figure 4). Thus, structures with varying degree of weakly resolved substructures can receive an identical EDIA$_m$. This reflects our opinion of only selecting well-supported structures with an automatic analysis. EDIA$_m$ is a measurement for graveness of structural inconsistency seen at least once in the set of atom scores, it does not strongly reflect the number of affected atoms leading to the respective EDIA$_m$. Structures with an EDIA$_m$ below 0.8 should be manually inspected before being used further. Structures with identical EDIA$_m$ can be discriminated in identifying the number and size of connected components with at least two heavy atoms that are well supported (EDIA of at least 0.8). The sum of all heavy atoms in such well-resolved substructures normalized by the number of the overall heavy atoms in the structure is called OPIA (overall percentage of well-resolved interconnected atoms). Examples for OPIA are given in the Results section, Figure 7.

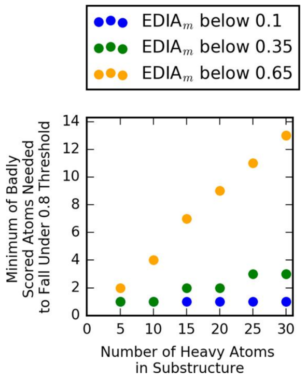


Figure 4. Development of EDIA$_m$ when increasing the number of problematic EDIA in the set of scores with the default score of 1.0. The y-axis shows the minimum number of scores with a value of 0.1 (blue), 0.35 (green), and 0.65 (orange) to let EDIA$_m$ drop below 0.8. In, e.g., a set of five atoms, one score of 0.35 and four of 1.0 result in an EDIA$_m$ below 0.8.

Numerical Stability. We examine the scoring range and numerical stability of EDIA in creating artificial test cases with various degrees of supporting electron density (SI 2.2). Each molecular fragment is then deflected up to 1 Å from its original position in either the x, z, or diagonal direction. Full experiments and results can be found in SI 2.3. EDIA shows a score development strongly coupled to the degree of displacement from the electron density. Only when unaccounted density surrounds the atom in the test case, EDIA fluctuates visibly. This is due to the different slopes in the positive and negative weighting curve as well as due to the oscillating number of grid points present in the EDIA calculation when moving the atom over the grid (see Figures S8 and S9).

Based on the artificial test cases, we aim to identify resolution independent scoring ranges for EDIA. Therefore, we inspect the EDIA values per constructed example at its initial position up to the displacement of half the grid space diagonal, which is the maximum positional uncertainty in the grid. As result, the following three EDIA quality intervals are identified:

- [0.8, 1.2]: Atom is well covered with electron density.
- [0.4, 0.8[: Atom shows minor inconsistencies with the electron density fit.
- [0.0, 0.4[: Atom shows substantial inconsistencies with the electron density fit.

The type of inconsistency can be identified by the integrated error analysis as introduced above. The three intervals are visually represented by blue, pink, and red atom colors following the color scheme in Figure 5.

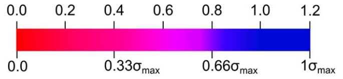


Figure 5. EDIA color scheme.

RESULTS

In a series of experiments we evaluate different aspects related to the performance of EDIA. We start in showing statistical results on EDIA$_m$ calculation on a large portion of the Protein Data Bank followed by the evaluation of EDIA on the Astex Diverse Set. We then present and discuss the relationship between EDIA$_m$ values and B factors as given in PDB structures. Subsequently, we compare EDIA and EDIA$_m$ values with RSCC and RSZD calculations. Finally, we analyze the robustness of EDIA$_m$ against structural shifts of the ligand molecule and compare the decrease of EDIA$_m$ values with the increase of RMSD values.

Automated Protein Structure Assessment. As starting point for a high quality structure data set, we screen all 32 844 protein−ligand complexes in the earlier defined high resolution PDB subset (SI 1). The run time per complex is on average 0.4 min on a Linux openSUSE 13.1 cluster equipped with Intel cores (2.2Ghz to 2.7 Ghz) and at least 8 GB RAM. The resulting EDIA$_m$ of the 45 113 ligands of interest[30] are given in Figure 6. The ligand EDIA$_m$ versus resolution plot can be found in SI 3.6. 77% of the ligands show an EDIA$_m$ of at least 0.8 and thus fit well in their electron density. Multiple examples are listed in Figure 7. The ligands in Figure 7a−c share an EDIA$_m$ of 0.2 due to badly supported atoms. The differing relative amount of well supported atoms is mirrored in the annotated OPIA as the overall percentage of well-resolved interconnected atoms. OPIA allows to discriminate between overall badly supported ligands (Figure 7a) and ligands with only a partially unsupported structure (Figure 7c). Ligands of the latter may be a starting point for subsequent structure optimization.

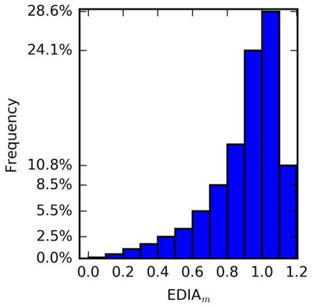
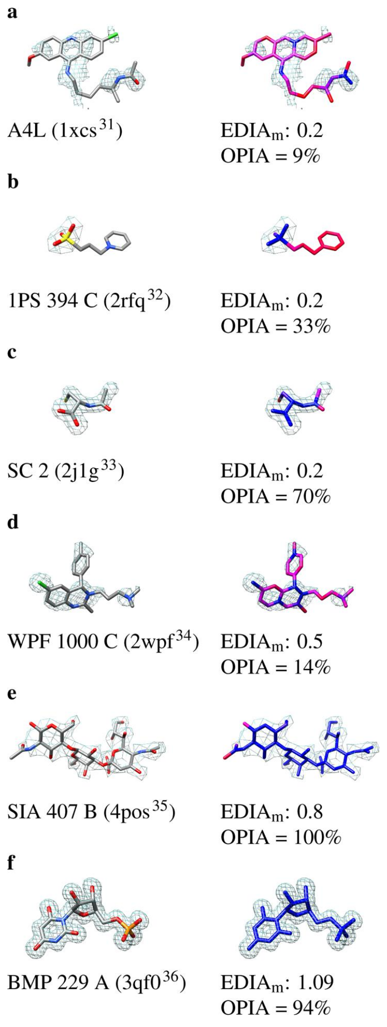


Figure 6. Distribution of all EDIA$_m$ of the 45 113 evaluated ligands in the high resolution PDB subset. 76.7% are well resolved with an EDIA$_m$ of at least 0.8.



A4L (1xcs[31]) EDIA$_m$: 0.2 OPIA = 9%

1PS 394 C (2rfq[32]) EDIA$_m$: 0.2 OPIA = 33%

SC 2 (2j1g[33]) EDIA$_m$: 0.2 OPIA = 70%

WPF 1000 C (2wpf[34]) EDIA$_m$: 0.5 OPIA = 14%

SIA 407 B (4pos[35]) EDIA$_m$: 0.8 OPIA = 100%

BMP 229 A (3qf0[36]) EDIA$_m$: 1.09 OPIA = 94%

Figure 7. Examples of ligands with differend EDIA$_m$ and the rounded percentage of atoms in good substructures (OPIA) in the high quality PDB subset. Parts a−c show a difference in OPIA while having the same EDIA$_m$ of 0.2. They are accompanied by three examples with an EDIA of 0.5, 0.8, and 1.09 to show the difference in density fit support. SC2, 1PS, and A4L have atoms with an occupancy below 1. SC2's sulfur has an EDIA value of 0.

Comparison with the Astex Diverse Set. The Astex Diverse Set consists of 85 protein−ligand complex structures with a resolution of at least 2.5 Å.[20] Hartshorn et al. calculated a density correlation between the *2fo* − *fc* and *fc* map[24] for structure selection and combined it with manual inspection of the electron density. We examine the Astex Diverse Set to evaluate the performance of EDIA$_m$ on a small, widely used data set. The distribution of all EDIA$_m$ values colored by resolution cutoff is displayed in Figure 8. The majority (81 of

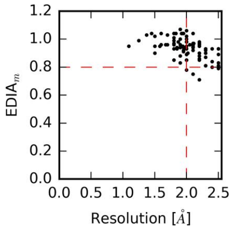


Figure 8. EDIA$_m$ versus resolution of all ligands in the Astex Diverse Set. Red lines mark the resolution cutoff at 2.0 Å and the EDIA$_m$ cutoff of 0.8. 47 ligands have an EDIA$_m$ of at least 0.8, and a resolution of 2.0 Å and smaller.

85) have an EDIA$_m$ value of at least 0.8. The lower EDIA$_m$ values in four cases express the need for manual inspection but can still be used in some validation scenarios. Biphenyl propanoic acid (1q4g,[37] Figure 9a) and butyl-dihydroimidazo-pyridazine-dione (1n2v,[38] Figure 9b) with atomic EDIA values around 0.27 both do not pass the automatic EDIA scan due to weakly supported carbon atoms. R-Roscovitine (1unl,[39] Figure 9c) and benzyl-benzyloxymethyl-isopropyl uracil (1jla,[40] Figure 9d) both have minimum atomic EDIA values around 0.49 accompanied by medium-supported benzole rings that let EDIA$_m$ drop below 0.8. We advise to also exclude structures with a resolution larger than 2.0 Å to avoid distorted electron density. Taking these two criteria, the historic Astex Diverse Set would be reduced to 47 structures.

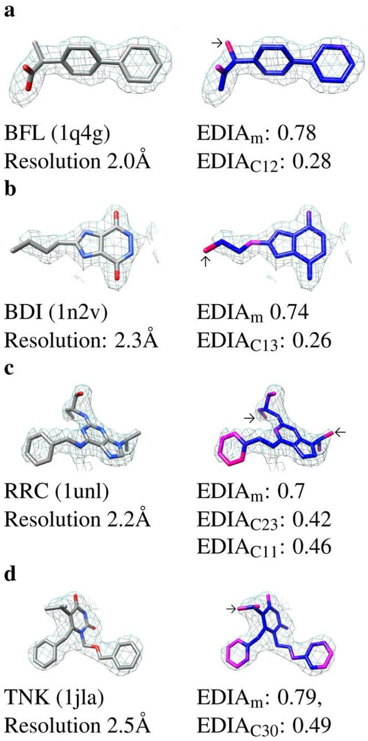


Figure 9. All ligands in the Astex Diverse Set with an EDIA$_m$ below 0.8. Minimal atomic EDIA scores are annotated and marked in the respective molecule. The $2fo - fc$ map is shown at 1σ.

B Factor Comparison. EDIA uses a resolution dependent B factor for the overall structure, disregarding the annotated atomic B factors. Thus, computing EDIA for a structure checks against unusual B factors and suggests an explanation for the problem with its integrated error analysis. In our search for example structures, we calculate the normed annotated B factors for every residue in the high resolution PDB data set (SI 1, 32 844 structures) and compare it against its EDIA$_m$ (SI Figure S13). The full analysis can be found in SI 3.2. In 16%, the annotated B factor disagrees with the EDIA$_m$. 5210 structures have at least one residue in which the annotated B factor is more than 175% the expected B factor while the EDIA$_m$ reports a well-supported residue (case **1**). In contrast, 36 structures have B factors of maximally 25% the expected B factor while the EDIA$_m$ reports a badly supported residue (case **2**). A case-**2** example is Isoleucin 126 A in the FIMH Lectin (5aap,[41] Figure 10a). The isoleucine side chain is present in four alternate conformations, all annotated with low B factors and an occupancy of 0.16−0.44. EDIA does not consider occupancy, and only reports the atom's qualitative fit with an automatic error analysis. Thus, EDIA detects missing electron density at, e.g., conformation A but leaves the interpretation up to the user. In contrast, residues 233 to 236 in chain A of Cytochrome P450 CYP 125 (3iw0,[42] Figure 10b) are case-**1** examples with B factors over 39 Å. While there is enough electron density to support atoms, it is also stretched out. Even though EDIA error analysis annotates numerous atoms with

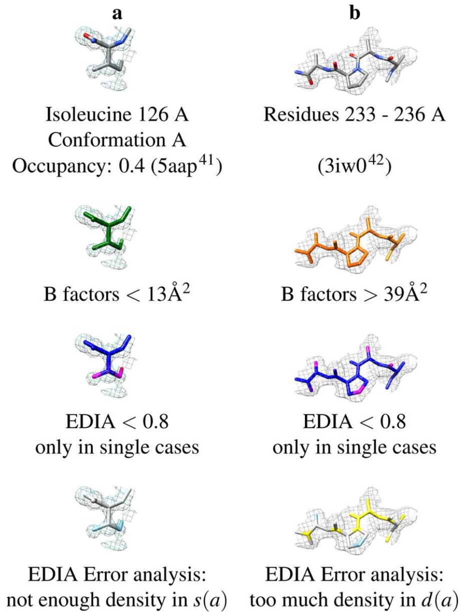


Figure 10. Examples in which B factor disagrees with EDIA quality categorization. B factor color gradient: red 50 Å^2; orange 40 Å^2; yellow 30 Å^2; light green 20 Å^2; dark green 5 Å^2. EDIA error analysis color code: white no error; yellow density in $d(a)$; light blue not enough density in $s(a)$. EDIA color scheme is shown in Figure 5.

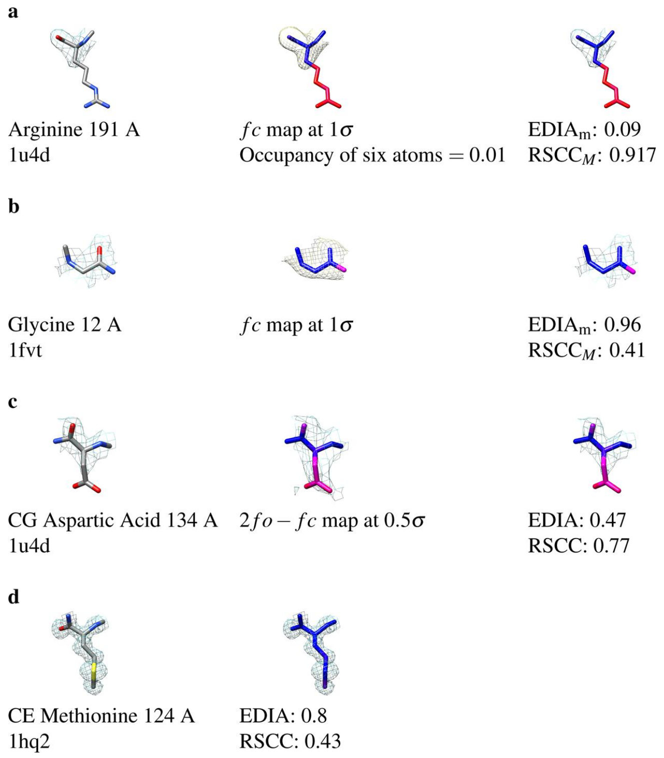


Figure 11. Four examples with Mapmans RSCC (RSCC$_M$) and EDIA$_m$ as well as atomic RSCC and EDIA scores are shown to display the differences between both scoring schemes. Each residue is colored in both its element and EDIA colors. The $2fo - fc$ map is shown at 1σ unless noted otherwise.

having unassigned electron density in their environment, EDIA still identifies most of the atoms as well-supported. Nevertheless, the EDIA values point attention to this regions for which a precise atom position is increasingly difficult to derive from electron density. Besides cases with only a few deviating values, our analysis also reports 211 structures in which at least 50% of the residues have a deviating B factor and EDIA$_m$ value. Of those, 205 fall into case-**1** for which an example was already presented. Case-**2** occurs in six structures: Xanthomonin (4nag[43]), Prion Protein (4ubz[44]), Cyclophilin (3rcg[45]), Z-DNA(4hig,[46] 2elg[47]), and the High-Potential Iron Sulfur Protein (3a39[48]). These six out of 201 proteins with a resolution of maximally 1 Å in the data set are very small. They also show highly compact electron density around each atom position, consequently resulting in EDIA$_m$ of approximately 0.7. We assume the shape of the electron density to be an artifact of the refinement procedure for structures with high resolution.

Correlation with RSCC. We correlate EDIA and EDIA$_m$ with the broadly adopted RSCC.[18] RSCC values for amino acids were calculated with an electron density sphere radius of 1.5 Å using Mapman.[49] For the evaluation, all 8283 residues closer than 10 Å to the ligand in the Iridium HT[21] structures are used. The EDIA$_m$ and RSCC agree in 84% of all cases in marking the residues as well-resolved with an overall correlation coefficient of 0.62 (see SI 3.3, Figure S15). 11% of the residues are differently categorized. The RSCC by Mapman is based on the precomputed fc map which considers annotated B factors and occupancies. In cases of low occupancies, weak density in the fc map may agree with weak density in the $2fo - fc$ map resulting in a high RSCC. On the contrary, EDIA$_m$ reports weak density in $2fo - fc$ map indicating poor electron density support for the structure. As an example, Arginine 191 A in CDC42 Kinase 1 (1u4d[50]) with an RSCC of 0.92, and an EDIA$_m$ of 0.09 is presented in Figure 11a. On the other hand, expected density may be not as voluminous as the observed density, resulting in a low RSCC. In some cases such as in Glycine 13 A in CDK2 (1fvt,[51] Figure 11b) EDIA$_m$ still accepts the residue as well supported due to the resolution dependent mean B factor in the EDIA computation.

Unfortunately, the Mapman implementation of the RSCC does not report RSCC values on the atomic level. We therefore used an in-house implementation based on the oversampled electron density grid of EDIA and a Gaussian shaped fc.[12] With a Pearson correlation coefficient of 0.86 on 66.009 data points, the RSCC and EDIA show a significant correlation (see Figure S16). Four examples are shown as representatives for deviating quality assessment (Figures 11c and d, S18). In the cases of Glutamate 509 A in Phosphodiesterase 4B (1xm6,[52] Figure S18a) and Aspartatic Acid 134 A in CDC42 Kinase 1 (1u4d,[50] Figure 11c), both residues have a voluminous electron density to support every atom in the residue but only at the intensity level of 0.5σ. Because RSCC does not explicitly include the intensity level, it scores the atoms as well supported. EDIA mirrors the intensity level of electron density at this position thus hinting toward lower density support. In the case of Methionine 124 A in the activated CDC42 Kinase (1hq2,[50] Figure 11d) and Cysteine 42 A in the Pyrophosphokinase (3ptb,[53] Figure S18b), the EDIA score declares the residue as well supported while the RSCC of the respective carbon marks it as badly resolved. In both cases, the electron density is slimmer than expected. RSCC relies on shape comparison integrating an *fc* map in the calculation. The shape and resulting effects of *fc* on the RSCC can be startling for the user. EDIA only uses the familiar $2fo - fc$ map with a weighting scheme to weight down fuzzy borders in its scoring calculation making EDIA values more intuitive and comprehensible.

Correlation with RSZD and RSZO. In the following, we compare EDIA_m to RSZD and RSZO as implemented in EDSTATS.[17] We compute the RSZD+, RSZD−, and RSZO of the backbone and side chain atoms as well as the EDIA_m for both atom sets in the binding pocket of the Iridium HT data set. The result is displayed in Figure S20. Tickle defines the threshold for the RSZD to be below −3 or above +3 for inaccurate structures.[17] In the case of RSZO, the local precision score should be at least 1σ to allow local interpretation of the electron density.[17] For about 85% in the case of RSZD and 86% in the case of RSZO of all atom sets, they and EDIA agree in marking the atom sets as well reproduced by the electron density. EDIA seems to be more sensitive with 11% of the atoms in the EDIA medium quality range while still in the [−3, + 3] range of RSZD values. 11% of all atom sets are also in the medium EDIA range while staying above RSZO's 1σ threshold. The backbone EDIA_m for Glutamate 241 I in Triosephosphate Isomerase (1ml1,[54] Figure 12a) and the side chain EDIA_m for Isoleucine 94 A in the HIV-1 Reverse Transcriptase (1c1b[55] Figure S21a) report only medium quality while the RSZD does not report any outliers and the RSZO reports a clear signal. The first case elucidates the difference between EDIA and RSZO. While RSZO reports the strength of the local signal, EDIA factors in the shape of the local data. Here, spacious electron density and high B factors are present. Serious disagreements between the two scoring schemes can be found in the EDIA range [0.0, 0.4[with unproblematic RSZD scores in 1% of all cases. RSZO and EDIA disagree in 1% of all cases as well. As example, Leucine 42 A in a Dihydroorotate Dehydrogenase (1d3h,[56] Figure 12b) and Lysine 140 A in Hemaglutinin (1hgg,[57] Figure S21d) are shown. They display a low side chain EDIA_m while having no significant RSZD peaks in both directions. While there is not enough electron density in the $2fo - fc$ map at a contour level of 1σ to pinpoint the exact position of the residue, the RSZD does not detect significant peaks in the $fo - fc$ map due to the high B factor of the substructures. In these two cases, the RSZO reports problematic local density though. The last case enlightens the different focus of RSZD and EDIA. While the RSZD is used by crystallographers to detect modeling inaccuracies, EDIA supports the modeler to distinguish reliable from less reliable atom positions. This ability is in parts carried out by RSZO but falls short in the case of voluminous density annotated with high B factors. EDIA safeguards the modeler against such unusual disorder beyond the range of the predetermined resolution dependent B factor. More examples can be found in SI 3.4

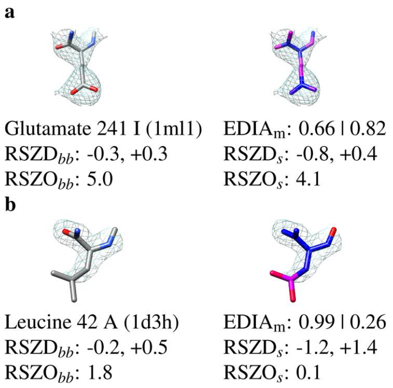
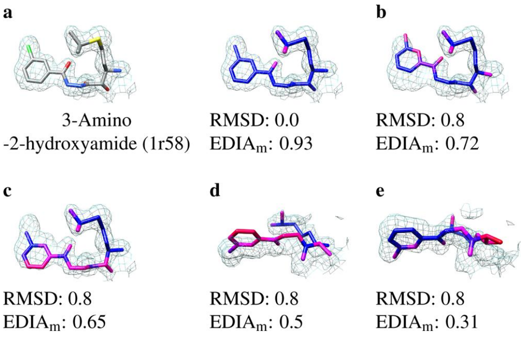


Figure 12. Selected residues with RSZD, RSZO, and EDIA scores. In each example, the RSZD and EDIA disagree in their quality assessment. Every example is shown in element and EDIA coloration. RSZD_{bb}, RSZO_{bb}:RSZD, or RSZO of the backbone atoms, RSZD_s, RSZO_s:RSZD, or RSZO of the side chain atoms. EDIA_m is listed in the ordering: EDIA_m for backbone, EDIA_m for side chain atoms. The $2fo - fc$ map is visualized with a contour level of 1σ. The $fo - fc$ map is shown above 3σ in green and below -3σ in red.

Evaluation of Spatial Displacement. Method validation in molecular modeling is heavily based on measuring spatial displacement between the experimentally resolved and computationally predicted molecular structures. It is measured with the root-mean-squared deviation (RMSD) in the range [0, ∞] with 0 signifying no deviation between the two structures. Depending on the application, an RMSD cutoff is selected to separate correctly from incorrectly placed structures. An RMSD of 2 Å is a frequently applied cutoff in the evaluation of docking experiments.[58] Since RMSD is based on absolute atom coordinates, it can not integrate local experimental uncertainties. In the search for better evaluation methods, Hawkins et al.[59] have considered the RSCC and RSR but found it lacking due to the not existing correlation with RMSD in the interval from 0 to 2 Å. Here, we examine this correlation between RMSD and EDIA_m in systematically exploring the conformational space of the respectively first ligand for Mc/Pc603 Fab-Phosphocholine Complex (2mcp[60]), Methionine Aminopeptidase 2 (1r58[61]), Phosphate Synthase (1of6[62]), Protein Kinase CHK1 (2br1[63]), and Beta-Xylanase (1fh9[64]) from the Iridium HT data set. The exact experiment is described in SI 3.5. Overall, at least 1764 poses per structure are systematically generated. The EDIA_m anticorrelates with a Pearson correlation coefficient of maximally −0.93 with the RMSD in all five cases (Table S5, Figure S22). The correlation plots reveal a sigmoid shape with an EDIA_m plateau up to an RMSD of 0.4 Å. From an RMSD of 1.5 Å on, the EDIA_m is in all cases below 0.2 marking the second plateau of the sigmoid shape. Figure 13 and S23 show several ligand poses of 3-amino-2-hydroxyamide (1r58) at an RMSD of 0.8 Å. The resulting EDIA_m, initially at 0.93, thus stretches from 0.72 to 0.31. While the RMSD measures the displacement of atoms from its model positions in any direction, EDIA_m describes the fit into the underlying electron density directly showing inconsistencies to the experimental data.

■ CONCLUSION

EDIA is a new method to quantify the electron density support of individual atoms in crystal structures with a resolution of up to 2 Å. In contrast to existing methods, EDIA does not rely on a structure to electron density conversion and takes the



3-Amino -2-hydroxyamide (1r58)

RMSD: 0.0 EDIA$_m$: 0.93

RMSD: 0.8 EDIA$_m$: 0.72

RMSD: 0.8 EDIA$_m$: 0.65

RMSD: 0.8 EDIA$_m$: 0.5

RMSD: 0.8 EDIA$_m$: 0.31

Figure 13. Examples for the correlation of EDIA$_m$ with RMSD of 3-amino-2-hydroxyamide (1r58). The initial ligand configuration is shown in element and EDIA coloring in part a. Four additional configurations of the ligand with identical RMSD are shown in EDIA coloring in parts b−e, annotated with RMSD and EDIA$_m$, and marked in blue in the correlation plot in Figure S22. (d and e) Rotated for a better understanding of the local electron density fit. They are shown in the original position in Figure S23. In all pictures, the $2fo - fc$ map is drawn at 1σ.

covalent structure of the molecules explicitly into account. The scores for a set of atoms such as a ligand, a residue or a functional group can be combined with the power mean. The resulting EDIA$_m$ value serves as an easily interpretable indicator score to discriminate between well and medium to badly resolved structural components. A rapid, objective, automatic, and reproducible analysis can thus be performed over any collection of atoms in a structure such as ligands, binding pockets, secondary structure elements, or complex interfaces. We examined EDIA scores on structures with different resolutions and the most frequent unit cell and grid spacing configurations. In this way, we were able to define generally applicable scoring intervals for EDIA values supporting noncrystallographers in interpreting the structure model.

We have shown numerous examples accompanied by comparisons to RSCC and RSZD illustrating the usefulness and practical advantages of EDIA$_m$. RSCC and RSZD both tolerate flexible atoms in integrating fc in their calculation. Although this makes sense from a crystallographic point of view, it make the analysis of spatial certainty for a modeler more complex. Instead, EDIA has a stricter concept of tolerable flexibility by using a mean B factor for the overall structure. We analyzed the behavior of EDIA$_m$ with respect to slight structural changes demonstrating the relationship to RMSD values. Also, EDIA was compared to B factors. It was shown, that the EDIA error analysis assists in B factor interpretation and that EDIA identifies highly unusual B factors. We calculated EDIA$_m$ values for a high resolution PDB subset of 32 844 structures classifying 77% of its ligands as well resolved. In automatically scoring electron density support in new structures in the PDB, it is now possible to create benchmark sets for modeling efforts of statistically significant size such as the Platinum data set.[65] With the help of the EDIA and its integrated error analysis, a guided tour through complexes of interest can be taken. Along with information from the PDB header, considering B factors, and occupancies as well as crystal contacts, EDIA values could help modelers and medicinal chemists to inspect a binding pocket of interest. It could also be integrated into the PDB quality assurance tools to detect poor substructure quality. Furthermore, EDIA also allows the comparison of a computed ligand pose to the original underlying experimental data. In the future, we will use EDIA for the evaluation of docking tools. In contrast to a state-of-the-art RMSD evaluation, EDIA$_m$ directly relates to the underlying electron density rather than to its interpreted model. Therefore, EDIA scores are able to support scientists along the whole process from structure elucidation to its use in molecular design.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00391.

Additional statistics, definition of used data sets, method and experiment description (PDF)

Additional files for computation: curves.ini: Density radii configuration file; edia_weighting_curve.ipynb, edia_weighting_curve.nb: Calculation of weighting curve; electron_density_radius_calcul ation.py: calculation of electron density radii; pdb_ids_high_resolution_PDB_set.list: PDB IDs of high resolution PDB subset; convert_mtz_to_ccp4.sh: MTZ-CCP4-Conversion file (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

ORCID

Agnes Meyder: 0000-0001-8519-5780
Eva Nittinger: 0000-0001-7231-7996
Matthias Rarey: 0000-0002-9553-6531

Notes

The authors declare no competing financial interest.

With this publication we release the EDIAscorer executable. The EDIAscorer executable is available free of charge for academic use as part of the AMD software bundle at http://www.zbh.uni-hamburg.de/edia. More information about the EDIAscorer can be found in SI3.6. Furthermore, EDIA values can be calculated and visually inspected with the ProteinsPlus server,[66] freely available at http://proteins.plus.

■ ACKNOWLEDGMENTS

Agnes Meyder and Eva Nittinger were partially funded by BMBF (031A183B) Project 47. The authors thank N. Friedrich, F. Flachsenberg, N. Pearce, and J. Freigang for helpful discussions. All data analysis was carried out with Python, IPython,[67] and Mathematica.[68] Chimera[69] 1.8 was used for the visualization of the structures.

■ REFERENCES

(1) Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* **2008**, *275*, 1−21.

(2) Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem., Int. Ed.* **2003**, *42*, 2718−2736.

(3) Read, R.; et al. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure* **2011**, *19*, 1395−1412.

(4) Deller, M. C.; Rupp, B. Models of protein-ligand crystal structures: trust, but verify. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 817−836.

(5) Cruickshank, D. W. J. Remarks about protein structure precision. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 583−601.

(6) Blundell, T.; Johnson, L. *Protein Crystallography*; Academic Press: London, 1976.

(7) Gore, S.; Velankar, S.; Kleywegt, G. J. Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2012**, *68*, 478−483.

(8) Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2012**, *68*, 352−367.

(9) Touw, W. G.; Vriend, G. BDB: databank of PDB files with consistent B-factors. *Protein Eng., Des. Sel.* **2014**, *27*, 457−462.

(10) Trueblood, K. N.; Bürgi, H. B.; Burzlaff, H.; Dunitz, J. D.; Gramaccioli, C. M.; Schulz, H. H.; Shmueli, U.; Abrahams, S. C. Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1996**, *52*, 770−781.

(11) Agarwal, R. C. A new least-squares refinement technique based on the fast Fourier transform algorithm. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *34*, 791−809.

(12) Diamond, R. A real-space refinement procedure for proteins. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1971**, *27*, 436−452.

(13) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 486−501.

(14) Molecular Operating Environment (MOE). http://www.chemcomp.com (January 14, 2016).

(15) Langer, G.; Cohen, S. X.; Lamzin, V. S.; Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **2008**, *3*, 1171−1179.

(16) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *47*, 110−119.

(17) Tickle, I. J. Statistical quality indicators for electron-density maps. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2012**, *68*, 454−467.

(18) Jones, T.; Kjeldgaard, M. Electron Density Map Interpretation. *Methods Enzymol.* **1997**, *277*, 173−208.

(19) Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An alternative method for the evaluation of docking performance: RSR vs RMSD. *J. Chem. Inf. Model.* **2008**, *48*, 1411−1422.

(20) 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*, 726−741.

(21) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discovery Today* **2012**, *17*, 1270−1281.

(22) Dunbar, J. B.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y.-N.; Kubish, G.; Gestwicki, J. E.; Stuckey, J. A.; Carlson, H. A. CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. *J. Chem. Inf. Model.* **2013**, *53*, 1842−1852.

(23) 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*, 1700−1716.

(24) Mooij, W. T. M.; Hartshorn, M. J.; Tickle, I. J.; Sharff, A. J.; Verdonk, M. L.; Jhoti, H. Automated Protein-Ligand Crystallography for Structure-Based Drug Design. *ChemMedChem* **2006**, *1*, 827−838.

(25) Nittinger, E.; Schneider, N.; Lange, G.; Rarey, M. Evidence of water molecules - a statistical evaluation of water molecules based on electron density. *J. Chem. Inf. Model.* **2015**, *55*, 771−783.

(26) Gutmanas, A.; et al. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* **2014**, *42*, D285−D291.

(27) Doublie, S., Ed. *Methods in Molecular Biology*; Humana Press: New Jersey, 2007; Vol. *2*, p 288.

(28) Ten Eyck, L. F. Efficient structure-factor calculation for large molecules by the fast Fourier transform. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1977**, *33*, 486−492.

(29) Jeudy, S.; Lartigue, A.; Claverie, J.-M.; Abergel, C. Dissecting the Unique Nucleotide Specificity of Mimivirus Nucleoside Diphosphate Kinase. *J. Virol.* **2009**, *83*, 7142−7150.

(30) Schomburg, K. T.; Bietz, S.; Briem, H.; Henzler, A. M.; Urbaczek, S.; Rarey, M. Facing the challenges of structure-based target prediction by inverse virtual screening. *J. Chem. Inf. Model.* **2014**, *54*, 1676−1686.

(31) Valls, N.; Steiner, R. A.; Wright, G.; Murshudov, G. N.; Subirana, J. A. Variable role of ions in two drug intercalation complexes of DNA. *JBIC, J. Biol. Inorg. Chem.* **2005**, *10*, 476−482.

(32) Chang, C.; Skarina, T.; Kagan, O.; Savchenko, A.; Edwards, A.; Joachimiak, A. Crystal structure of 3-HSA hydroxylase, oxygenase from Rhodococcus sp. RHA1. *PDB* **2007**; DOI: 10.2210/pdb2rfq/pdb.

(33) Garlatti, V.; Belloy, N.; Martin, L.; Lacroix, M.; Matsushita, M.; Endo, Y.; Fujita, T.; Fontecilla-Camps, J. C.; Gaboriaud, C.; Arlaud, J.; Thielens, N. M. Structural insights into the innate immune recognition specificities of L- and H-ficolins. *EMBO J.* **2007**, *26*, 623−633.

(34) Patterson, S.; Alphey, M. S.; Jones, D. C.; Shanks, E. J.; Street, I. P.; Frearson, J. A.; Wyatt, P. G.; Gilbert, I. H.; Fairlamb, A. H. Dihydroquinazolines as a novel class of Trypanosoma brucei trypanothione reductase inhibitors: Discovery, synthesis, and characterization of their binding mode by protein crystallography. *J. Med. Chem.* **2011**, *54*, 6514−6530.

(35) Khan, Z. M.; Liu, Y.; Neu, U.; Gilbert, M.; Ehlers, B.; Feizi, T.; Stehle, T. Crystallographic and Glycan Microarray Analysis of Human Polyomavirus 9 VP1 Identifies N-Glycolyl Neuraminic Acid as a Receptor Candidate. *J. Virol.* **2014**, *88*, 6100−6111.

(36) Desai, B. J.; Wood, B. M. K.; Fedorov, A. A.; Fedorov, E. V.; Goryanova, B.; Amyes, T. L.; Richard, J. P.; Almo, S. C.; Gerlt, J. A. Conformational changes in orotidine 5′-monophosphate decarboxylase: A structure-based explanation for how the 5-phosphate group activates the enzyme. *Biochemistry* **2012**, *51*, 8665−8678.

(37) Gupta, K.; Selinsky, B. S.; Kaub, C. J.; Katz, A. K.; Loll, P. J. The 2.0 Å Resolution Crystal Structure of Prostaglandin H 2 Synthase-1: Structural Insights into an Unusual Peroxidase. *J. Mol. Biol.* **2004**, *335*, 503−518.

(38) Brenk, R.; Naerum, L.; Grädler, U.; Gerber, H. D.; Garcia, G. A.; Reuter, K.; Stubbs, M. T.; Klebe, G. Virtual screening for submicromolar leads of tRNA-guanine transglycosylase based on a new unexpected binding mode detected by crystal structure analysis. *J. Med. Chem.* **2003**, *46*, 1133−1143.

(39) Mapelli, M.; Massimiliano, L.; Crovace, C.; Seeliger, M. A.; Tsai, L. H.; Meijer, L.; Musacchio, A. Mechanism of CDK5/p25 binding by CDK inhibitors. *J. Med. Chem.* **2005**, *48*, 671−679.

(40) Ren, J.; Nichols, C.; Bird, L.; Chamberlain, P.; Weaver, K.; Short, S.; Stuart, D. I.; Stammers, D. K. Structural mechanisms of drug resistance for mutations at codons 181 and 188 in HIV-1 reverse transcriptase and the improved resilience of second generation non-nucleoside inhibitors. *J. Mol. Biol.* **2001**, *312*, 795−805.

(41) De Ruyck, J.; Lensink, M. F.; Bouckaert, J. Structures of C-mannosylated anti-adhesives bound to the type 1 fimbrial FimH adhesin. *IUCrJ* **2016**, *3*, 163−167.

(42) Tsurumura, T.; Qiu, H.; Yoshida, T.; Tsumori, Y.; Hatakeyama, D.; Kuzuhara, T.; Tsuge, H. Conformational polymorphism of m7GTP in crystal structure of the PB2 middle domain from human influenza a virus. *PLoS One* **2013**, *8*, e82020.

(43) Hegemann, J. D.; Zimmermann, M.; Zhu, S.; Steuber, H.; Harms, K.; Xie, X.; Marahiel, M. A. Xanthomonins I-III: A new class of lasso peptides with a seven-residue macrolactam ring. *Angew. Chem., Int. Ed.* **2014**, *53*, 2230−2234.

(44) Yu, L.; Lee, S.-J.; Yee, V. C. Crystal Structures of Polymorphic Prion Protein *β*1 Peptides Reveal Variable Steric Zipper Conformations. *Biochemistry* **2015**, *54*, 3640−3648.

(45) Ahmed-Belkacem, A.; Colliandre, L.; Ahnou, N.; Nevers, Q.; Gelin, M.; Bessin, Y.; Brillet, R.; Cala, O.; Douguet, D.; Bourguet, W.; Krimm, I.; Pawlotsky, J.-M.; Guichou, J.-F. Fragment-based discovery of a new family of non-peptidic small-molecule cyclophilin inhibitors with potent antiviral activities. *Nat. Commun.* **2016**, *7*, 12777.

(46) Drozdzal, P.; Gilski, M.; Kierzek, R.; Lomozik, L.; Jaskolski, M. Ultrahigh-resolution crystal structures of Z-DNA in complex with Mn^{2+} and Zn^{2+} ions. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2013**, *69*, 1180−1190.

(47) Ohishi, H.; Tozuka, Y.; Da-Yang, Z.; Ishida, T.; Nakatani, K. The rare crystallographic structure of d(CGCGCG)2: The natural spermidine molecule bound to the minor groove of left-handed Z-DNA d(CGCGCG)2 at 10°C. *Biochem. Biophys. Res. Commun.* **2007**, *358*, 24−28.

(48) Takeda, K.; Kusumoto, K.; Hirano, Y.; Miki, K. Detailed assessment of X-ray induced structural perturbation in a crystalline state protein. *J. Struct. Biol.* **2010**, *169*, 135−144.

(49) Kleywegt, G. J.; Jones, T. A. xdlMAPMAN and xdlDATAMAN − Programs for Reformatting, Analysis and Manipulation of Biomacromolecular Electron-Density Maps and Reflection Data Sets. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1996**, *52*, 826−828.

(50) Lougheed, J. C.; Chen, R. H.; Mak, P.; Stout, T. J. Crystal structures of the phosphorylated and unphosphorylated kinase domains of the Cdc42-associated tyrosine kinase ACK1. *J. Biol. Chem.* **2004**, *279*, 44039−44045.

(51) Davis, S. T.; et al. Prevention of chemotherapy-induced alopecia in rats by CDK inhibitors. *Science* **2001**, *291*, 134−137.

(52) Card, G. L.; et al. Structural basis for the activity of drugs that inhibit phosphodiesterases. *Structure* **2004**, *12*, 2233−2247.

(53) Blaszczyk, J.; Li, Y.; Shi, G.; Yan, H.; Ji, X. Dynamic roles of arginine residues 82 and 92 of Escherichia coli 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase: Crystallographic studies. *Biochemistry* **2003**, *42*, 1573−1580.

(54) Thanki, N.; Zeelen, J. P.; Mathieu, M.; Jaenicke, R.; Abagyan, R. A.; Wierenga, R. K.; Schliebs, W. Protein Eng. with monomeric triosephosphate isomerase (monoTIM): the modelling and structure verification of a seven-residue loop. *Protein Eng., Des. Sel.* **1997**, *10*, 159−167.

(55) Hopkins, A. L.; Ren, J.; Tanaka, H.; Baba, M.; Okamato, M.; Stuart, D. I.; Stammers, D. K. Design of MKC-442 (emivirine) analogues with improved activity against drug-resistant HIV mutants. *J. Med. Chem.* **1999**, *42*, 4500−4505.

(56) Liu, S.; Neidhardt, E. A.; Grossman, T. H.; Ocain, T.; Clardy, J. Structures of human dihydroorotate dehydrogenase in complex with antiproliferative agents. *Structure* **2000**, *8*, 25−33.

(57) Sauter, N. K.; Hanson, J. E.; Glick, G. D.; Brown, J. H.; Crowther, R. L.; Park, S. J.; Skehel, J. J.; Wiley, D. C. Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography. *Biochemistry* **1992**, *31*, 9609−96021.

(58) Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B.; Stuckey, J. A.; Carlson, H. A. CSAR benchmark exercise 2011−2012: evaluation of results from docking and relative ranking of blinded congeneric series. *J. Chem. Inf. Model.* **2013**, *53*, 1853−1870.

(59) Hawkins, P. C. D.; Kelley, B. P.; Warren, G. L. The application of statistical methods to cognate docking: A path forward? *J. Chem. Inf. Model.* **2014**, *54*, 1339−1355.

(60) Padlan, E.; Cohen, G.; Davies, D. Refined Crystal Structure of the Mc/Pc603 Fab-Phosphocholine Complex at 3.1 Angstroms Resolution. *PDB* **1984**; DOI: 10.2210/pdb2mcp/pdb.

(61) Sheppard, G. S.; Wang, J.; Kawai, M.; BaMaung, N. Y.; Craig, R. A.; Erickson, S. A.; Lynch, L.; Patel, J.; Yang, F.; Searle, X. B.; Lou, P.; Park, C.; Kim, K. H.; Henkin, J.; Lesniewski, R. 3-Amino-2-hydroxyamides and related compounds as inhibitors of methionine aminopeptidase-2. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 865−868.

(62) Koenig, V.; Pfeil, A.; Heinrich, G.; Braus, G.; Schneider, T. Crystal Structure of the Double Complex of the Tyrosine Sensitive Dahp Synthase from Yeast. *PDB* **2004**; DOI: 10.2210/pdb1of6/pdb.

(63) Foloppe, N.; Fisher, L. M.; Howes, R.; Kierstan, P.; Potter, A.; Robertson, A. G. S.; Surgenor, A. E. Structure-based design of novel Chk1 inhibitors: Insights into hydrogen bonding and protein-ligand affinity. *J. Med. Chem.* **2005**, *48*, 4332−4345.

(64) Notenboom, V.; Williams, S. J.; Hoos, R.; Withers, S. G.; Rose, D. R. Detailed Structural Analysis of Glycosidase/Inhibitor Interactions: Complexes of Cex from Cellulomonas fimi with Xylobiose-Derived Aza-Sugars. *Biochemistry* **2000**, *39*, 11553−11563.

(65) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 529−539.

(66) Fährrolfes, R.; Bietz, S.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; Volkamer, A.; Rarey, M. ProteinsPlus: a web portal for structure analysis of macromolecules. *Nucleic Acids Res.* **2017**, *45*, W337.

(67) Pérez, F.; Granger, B. E. IPython: A system for interactive scientific computing. *Comput. Sci. Eng.* **2007**, *9*, 21−29.

(68) Wolfram Research Inc. *Mathematica*, Version 9.0; 2012.

(69) Yang, Z.; Lasker, K.; Schneidman-Duhovny, D.; Webb, B.; Huang, C. C.; Pettersen, E. F.; Goddard, T. D.; Meng, E. C.; Sali, A.; Ferrin, T. E. UCSF Chimera, MODELLER, and IMP: An integrated modeling system. *J. Struct. Biol.* **2012**, *179*, 269−278.

*NAOMI*nova: Interactive Geometric Analysis of Noncovalent Interactions in Macromolecular Structures.

[D4] Inhester, T.; **Nittinger, E.**; Sommer, K.; Schmidt, P.; Bietz, S.; Rarey, M. *NAOMI*nova: Interactive Geometric Analysis of Noncovalent Interactions in Macromolecular Structures. J. Chem. Inf. Model. 2017, 57 (9): 2132-2142.

`http://pubs.acs.org/articlesonrequest/AOR-4TsN3Sfez6AjEHcHZUzt`

Journal of
CHEMICAL INFORMATION
AND MODELING

Article

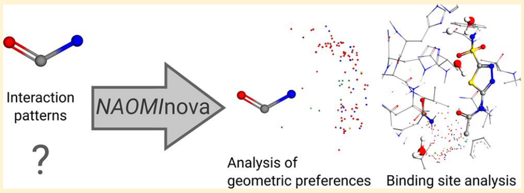
pubs.acs.org/jcim

*NAOMI*nova: Interactive Geometric Analysis of Noncovalent Interactions in Macromolecular Structures

Therese Inhester,[‡] Eva Nittinger,[‡] Kai Sommer, Pascal Schmidt, Stefan Bietz, and Matthias Rarey[*]

ZBH - Center for Bioinformatics, Universität Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

Supporting Information



ABSTRACT: Noncovalent interactions play an important role in macromolecular complexes. The assessment of molecular interactions is often based on knowledge derived from statistics on structural data. Within the last years, the available data in the Brookhaven Protein Data Bank has increased dramatically, quantitatively as well as qualitatively. This development allows the derivation of enhanced interaction models and motivates new ways of data analysis. Here, we present a method to facilitate the analysis of noncovalent interactions enabling detailed insights into the nature of molecular interactions. The method is integrated into a highly variable framework enabling the adaption to user-specific requirements. *NAOMI*nova, the user interface for our method, allows the generation of specific statistics with respect to the chemical environment of substructures. The substructures as well as the analyzed set of protein structures can be chosen arbitrarily. Although *NAOMI*nova was primarily made for data exploration in protein–ligand crystal structures, it can be used in combination with any structure collection, for example, analysis of a carbonyl in the neighborhood of an aromatic ring on a set of structures resulting from a MD simulation. Additionally, a filter for different atom attributes can be applied including the experimental support by electron density for single atoms. In this publication, we present the underlying algorithmic techniques of our method and show application examples that demonstrate *NAOMI*nova's ability to support individual analysis of noncovalent interactions in protein structures. *NAOMI*nova is available at http://www.zbh.uni-hamburg.de/naominova.

INTRODUCTION

Noncovalent interactions play a major role in the selectivity and affinity within a protein, in protein–protein interactions, as well as between small molecules and proteins. A thorough understanding of their geometrical preferences within their chemical context is mandatory for structure-based design projects from medicinal chemistry to biocatalysis and pesticide development. The assessment of molecular interactions is often based on knowledge derived from statistics on structural data.[1,2]

The main public resource for protein structures, the Brookhaven Protein Data Bank (PDB),[3] is constantly growing quantitatively as well as qualitatively. Because the focus on data quality has increased over the last years, structure factors are mandatory for new resolved structures deposited in the PDB since 2008. In February 2017, structure factors for almost 60 000 PDB files were available from the PDBe Web site (Protein data bank in Europe, http://www.ebi.ac.uk/pdbe/). This way, the verification of the experimental support for a structure is possible. Given this wealth of data, tools are needed that are able to analyze and visualize geometrical distributions of noncovalent interactions in large sets of protein complexes taking the quality of atomic positions into account.

In 1997, the tool IsoStar[4] was published, which is, to our knowledge, until today the only tool that can be used to analyze the geometric distribution of specific atoms or functional groups around a central molecular structure. IsoStar presents data derived from the Cambridge Structural Database (CSD)[5] and the PDB. For the used PDB structures, the main quality criterion used is a resolution of below 2.0 Å. A predefined set of frequently occurring functional groups was used to generate the data for each of its combinations. The data are stored in separate files, which have to be loaded separately into the program. User-defined substructures can only be added from the CSD data set using a combination of two other tools, IsoGen and ConQuest.[6] Moreover, the IsoStar user interface provides only a few possibilities to further analyze the collected data in detail, for example, selecting only atoms from a specific amino acid or only looking at inter/intramolecular interactions.

In addition to IsoStar, the tool SuperStar[7] uses the data from IsoStar to highlight hot spots within a protein–ligand interface, indicating the likelihood of a specific functional group to occur in this region. Also in this context, advanced filters to adapt the presented data to specific needs are missing.

Here, we present *NAOMI*nova, a new way for interactively analyzing geometric preferences of noncovalent interactions around user-defined substructures on any user-selected data. *NAOMI*nova stores all relevant data from a collection of protein structures in an SQLite database. This data can be used to analyze the distribution of interaction atoms in the vicinity of

Received: May 22, 2017
Published: September 11, 2017

DOI: 10.1021/acs.jcim.7b00291
J. Chem. Inf. Model. 2017, 57, 2132−2142

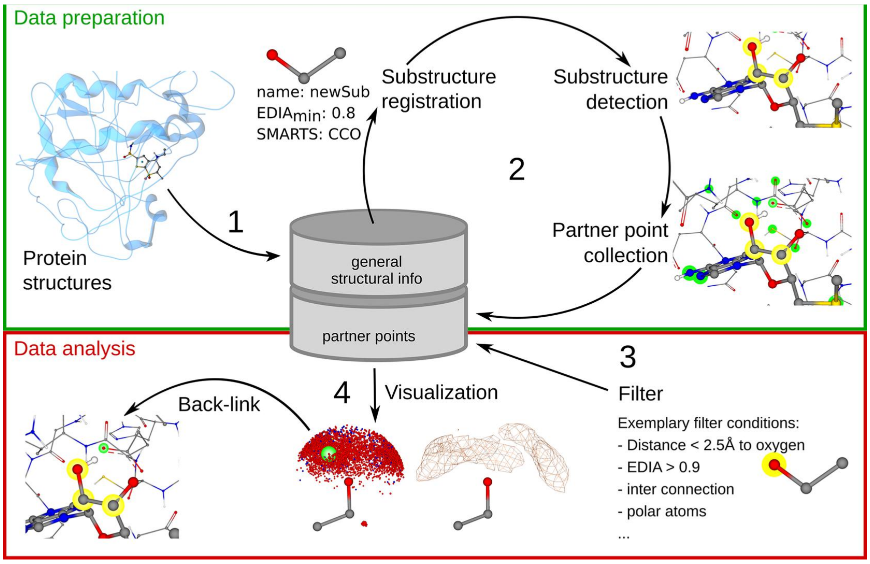


Figure 1. Schematic overview of the *NAOMI*nova method. The two data preparation steps are highlighted in green. The data analysis steps are highlighted in red. (1) A database is compiled out of protein structures. (2) New substructures are added to a database. This process includes registrations and detection of the substructures and the collection of partner points. (3) A set of partner points for a given substructure can be generated using the filtering process. (4) A set created by filtering can be visualized. The back-link functionality can be used to inspect the original structure for every partner point.

user-defined substructures. Different filters allow precise tailoring of the presented interacting atoms according to the demand of a user. Moreover, the experimental support of individual atoms can be taken into account using the EDIA value (electron density of individual atoms).[8,9] Thus, the analyzed data set can be separated into those with good electron density and those that have to be analyzed more carefully. As an example, the following questions can conveniently be answered using *NAOMI*nova: Do oxygens in ligands and waters have different preferred geometrical parameters when involved in a hydrogen bond to an amide? What are the most frequent interaction directions of a specific substructure that occurs in a ligand? What are the resulting geometries of a specific interaction in all structures resulting from my latest MD simulation?

With these features, *NAOMI*nova goes beyond the IsoStar approach in various aspects. First, any collection of protein structures can be analyzed. This can be beneficial if, for example, only structures derived from an in-house data set should be analyzed or if statistical distributions within all structures of a MD simulation should be conducted. Second, substructures for which distributions of interacting atoms are calculated can be defined by a user. Third, the calculated data in *NAOMI*nova are stored in one SQLite database. Thus, all data can be loaded with one file and different data sets can be created from one individual database. Moreover, the fast query operations provided by SQLite make interactive filtering by various filter criteria possible. Fourth, *NAOMI*nova uses a much more exact quality criterion than the overall resolution of a structure.

In this publication, we present the data handling and give a general overview of *NAOMI*nova. Finally, we show examples for geometric analyses with *NAOMI*nova and its potential application in drug development tasks. The graphical user interface (GUI) connected to *NAOMI*nova is presented and briefly explained in the Supporting Information (see the section entitled *Graphical User Interface*).

■ METHODS

Overview. The methodology behind *NAOMI*nova comprises four steps (schematically illustrated in Figure 1). First of all, a database is compiled out of a collection of protein structures. During this step, all relevant data, including the structural data of a protein and its ligands, are stored in an SQLite database. This process is explained in more detail in the Database Construction section. Afterward, molecular substructures can be registered and added to the database. Herein, all occurrences of the defined substructures are detected in the protein structures and their ligands. Interacting atoms in the vicinity of these occurrences (in the following called partner points) are collected and also stored in the database. This part is explained in the Substructure Registration and Substructure Detection and Data Collection sections. A database created in this process is stored as one single file and can be subjected to many analyses. Protein structures can be added to an existing database, and also substructures can be included or deleted. Hence, the first two steps can be seen as preparation steps that have to be performed only occasionally.

Using a compiled database, a set of partner points in the surrounding of a registered substructure can be filtered according to the demands of a user. The filter comprises different attributes of the partner point, for example, the element type, its connection to the substructure (intra or inter), or its distance to the substructure. This step is explained in the Filtering section. Finally, the filtered set of partner points can be visualized and the distribution can be analyzed geometrically.

DOI: 10.1021/acs.jcim.7b00291
J. Chem. Inf. Model. 2017, 57, 2132−2142

For each visualized partner point, a link to its original structure is stored, called the back-link. Hence, for each partner point, the original structure can be visualized in order to deduce reasons for specific structural characteristics. The visualization is explained in the Data Visualization section.

Database Construction. *NAOMI*nova allows the analysis of potential interaction partners of custom-defined chemical substructures. The data required for the detection and visualization of interaction partners are stored in an SQLite database. As a first step, all protein structures are processed to compile a database file. In this procedure, structural information on a given input file is interpreted using the NAOMI software library[10] and stored in different database tables. The same database technology for proteins and small molecules has been used previously[11−13] and will therefore only be described briefly here. The main purpose of the used tables is to store the relevant information on protein−ligand complexes in a compressed way, efficiently providing the necessary structural information upon request. Conceptually, small molecules and proteins are stored in different tables of the database. This separation is done because different attributes are required for the two different molecule classes, for example, chain ID or the amino acid type. Herein, a small molecule is defined as a chemical component from an input file consisting of less than six residues. Hence, this category comprises water molecules, metal ions, short peptides, and ligands. For small molecules, two different tables exist. The topology of small molecules is stored in one table using a textual representation. The 3D coordinates are stored in a second table. This way, the topology of a small molecule occurring several times in a data set has to be stored only once. On the opposite, the coordinates are stored for each small molecule. For example, the topology of a water molecules is only stored once in the database; however, the exact coordinates for each water molecule are stored each time in combination with a reference key to the respective topology.

For proteins, the topology and the 3D coordinates of individual amino acids are stored in two different tables. Here, the same concept as that used for small molecules is applied. The topology of an amino acid is only stored once, whereas the 3D coordinates are stored for every amino acid. A third table stores the connection of individual amino acids using foreign keys. Protein−ligand complexes are represented by only one table in our database concept. Herein, foreign keys mapping to small molecules as well as to proteins are stored pairwise. A ligand is here defined as a small molecule that is not water and not metal. Before entered into the database, structures are preprocessed with Protoss to decide on tautomeric forms and protonation states and to optimize the hydrogen bond network.[14] The optimization also includes flipping ambiguous conformations of histidine, glutamine, and asparagine that cannot be distinguished from the electron density itself. Additionally, the EDIA value for each atom is calculated and stored in a separate table. EDIA values represent the experimental electron density support of individual atoms. They can only be computed for structures with a resolution below 2.5 Å and if an 2fo-fc electron density map is available.[8,9] The EDIA ranges between 0.0 and 1.2. Atoms having an EDIA value above 0.8 are considered well supported by experimental data.[8,9]

Substructure Registration. During the substructure registration process, substructures can be defined by a user and are checked for compatibility. Overall, every substructure is associated with four attributes: (1) a 3D template molecule, (2) a molecular pattern (SMARTS[15]), (3) a unique name, and (4) an EDIA$_{\mathrm{min}}$. In the following, these attributes are explained in more detail.

3D Template Molecule. Each substructure requires a 3D template molecule in order to ensure the conformational identity of the detected hits. There are two different possibilities how the molecular structure of this template can be defined. First of all, a small molecule with 3D coordinates can be provided by the user. Alternatively, the topology of the template molecules can be defined by a SMILES[16] pattern. In this case, the 3D coordinates are calculated using UNICON.[17] From this molecular structure, atoms can be selected that define the geometry of the substructure.

Molecular Patterns. The molecular pattern specified with the SMARTS language defines the exact molecular substructure that will be searched in the database. The SMARTS pattern can be autogenerated by selecting heavy atoms in the 3D template molecule. Afterward, the pattern can be adapted manually. Conceptually, we differentiate between two different parts using SMARTS recursions—a fragment part and a molecular context. A graphical explanation of the SMARTS concept used in *NAOMI*nova is shown in Figure 2a. The fragment part describes a molecular fragment. This fragment has to match the selected atoms in the 3D template molecule. All matching atoms detected in the protein or ligand later will be superimposed onto these template atoms. Thus, every atom described within this part of the expression corresponds to exactly one atom in the molecular fragment. Atomic ambiguities such as the logical OR (SMARTS symbol: “,”) or a bond of any type (SMARTS symbol: “∼”) are in principle allowed here. However, these ambiguities may lead to hits that differ in their molecular structure, for example, different bond

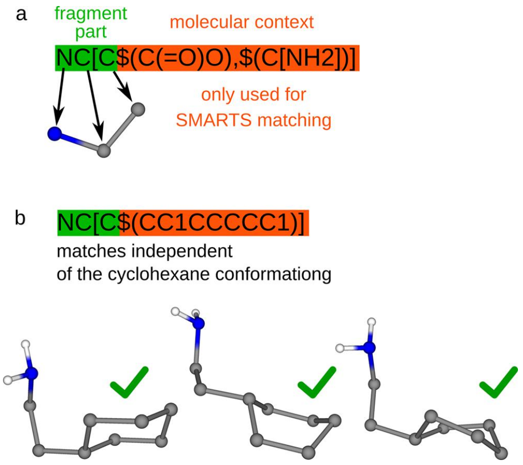


Figure 2. Schematic explanation of the SMARTS concept used in *NAOMI*nova. The fragment part is depicted in green. The molecular context of the SMARTS is depicted in orange. (a) Example of a SMARTS pattern. Each atom in the fragment part corresponds to exactly one atom in the 3D molecular template, indicated by black arrows. (b) The molecular context is not used for superimposing. Therefore, the depicted SMARTS pattern matches all three molecules while geometric matching is limited to the Ethylamine fragment.

lengths or angles. In the data collection step, only small geometric deviations between the template atoms and the detected matching atoms will be accepted (see the Substructure Detection and Data Collection section for detailed information). Moreover, the described molecular fragment must have at least two atoms to avoid point symmetries.

Each atom of the substructure can be further specified using the recursive notation of the SMARTS language (molecular context). The molecular context is not used for superimposing but only for the detection of substructures in a SMARTS matching procedure. The further specification of each atom with recursive SMARTS allows the definition of an unambiguous context, independent of the exact conformation of these atoms. Figure 2b shows an example of such a scenario. Here, a substructure with cylohexane as its molecular context is defined. Because cyclohexane can have different conformations, it should be specified in the molecular context. This way, hits with any cyclohexane conformation will be found in the data set.

Unique Name. For each substructure, a unique name has to be provided. This name is used as a unique key in the database.

$EDIA_{min}$. A lower bound for the EDIA value ($EDIA_{min}$) has to be defined for each substructure. Detected hits of the substructures in a protein and in ligands are only processed if their combined EDIA value ($EDIA_m$[8,9]) is above this lower bound. Additionally, partner points in the vicinity of a detected match are only used if they fulfill this lower bound, as well. A short explanation of the $EDIA_m$ calculation is provided in the Supporting Information (see the *$EDIA_m$ Calculation Details* section).

During the registration process, the validity of the attributes is checked, including the uniqueness of the name and the correctness of the SMARTS pattern. Afterward, all attributes are stored in a specific table of the database using the unique name as a key.

Substructure Detection and Data Collection. For data aggregation, each SMARTS pattern is searched in all proteins and ligands found in the protein structure collection employing a standard substructure matching algorithm.[18] In general, we are following the procedure described in Nittinger et al.[19] For all hits fulfilling the $EDIA_{min}$ criteria, the RMSD between the matching atoms and the corresponding atoms in the 3D template molecule is calculated. If the RMSD is below 0.2 Å, the matching atoms are transformed onto the template atoms. The low cutoff value of the RMSD was chosen to ensure a superimposition with low deviation necessary to derive meaningful data. For subsequent data collection, partner points in the vicinity of the matching atoms are detected. Herein, a partner point has to fulfill four criteria:

- It has a maximal distance of 4.5 Å to at least one matching atom of the substructure.
- Its element type is oxygen, nitrogen, sulfur, a halogen, or a metal.
- The EDIA of the partner point has to be larger than $EDIA_{min}$ of the current substructure.
- It is not connected to the matching atoms via four bonds or less.

These parameters were chosen in order to only hit interacting atoms around a central substructure. Apart from the EDIA criterion, these initial parameters are not user-adjustable and can only be further restricted by the subsequent filtering process. All partner points are transformed into the reference coordinate system of the template molecule. Each partner point is then stored in the database alongside several attributes, for example, its element type, its EDIA value, and its coordinates. The corresponding atom in the original protein or ligand structure is stored for each partner point in order to be able to provide the original structure for each data point. Moreover, the substructures a partner point is part of are stored using all currently registered substructures from the database.

Due to possible symmetries within the substructures, three different cases of symmetry have to be considered during the data collection step: (i) point symmetry, (ii) rotational symmetry, and (iii) substructure symmetry. A point symmetry was excluded by not allowing substructures containing only a single atom. Rotational symmetry can occur if, for example, a substructure contains two atoms only. In these cases, the detected partner points are randomly distributed on a circle around the symmetry axis. Substructure symmetries may occur if planar substructures are used, for example, a benzene ring. Here, the SMARTS pattern would match the same set of atoms several times. In these cases, all detected substructure hits are stored individually. Additionally, the information about the number of hits is stored in the database. Thus, in the data visualization, the user can decide if all symmetry hits should be displayed or only an arbitrarily picked first one.

Filtering. For the filter process, only the definition of a central substructure is mandatory. All other filter criteria for partner points, as element type, location (e.g., protein, ligand), or distance, are optional. A list of all possible filter criteria and their values are shown in the Supporting Information (Table S1). All resulting partner points are collected in a set, which can be visualized and further analyzed concerning its spatial distribution as well as its structural origins.

Data Visualization. For data visualization, a GUI has been developed. An overview of this interface and screenshots are provided in the Supporting Information (see the *Graphical User Interface* section). Briefly, the sets created by filtering can be visualized either with the used 3D template of the substructure (set view) or in a pocket of interest (pocket view). In both visualizations, the partner points in a set can be displayed either as spheres or as a density grid. The density grid is calculated by first placing a grid with a spacing of 0.4 Å onto all partner points. Using trilinear interpolation, the density of each grid point is determined. The exact calculation of the density values for each grid point is explained in the Supporting Information (see the *Density Grid Calculation* section).

By selecting a partner point in any of the two visualizations, the original structure is shown in a separate tab using the back-link functionality. Herein, the atoms in the original structure corresponding to the partner point and the substructure are highlighted, and the broader chemical environment can be analyzed.

Moreover, histograms for distance and angle distributions can be generated. Besides absolute counts for each bin, volume-normalized values are shown for distances and spherical angles. The normalization is explained in detail in the Supporting Information (see the *Histogram Normalization* section).

■ RESULTS AND DISCUSSION

In the following, performance analyses of the main functionalities are presented: (i) database construction, (ii) substructure detection and data collection, and (iii) filtering. In the second part, example applications are shown demonstrating *NAOMI-*

nova's ability to spatially analyze interactions and to guide molecular design projects.

Performance. First of all, a data set containing all PDB structures with a resolution better than 2.5 Å and with available electron density was created (downloaded November 2016 from PDB and PDBe). This data set, denoted in the following as "$PDB_{2.5}$", contains 56 807 protein−ligand structures. For $PDB_{2.5}$, a *NAOMI*nova database was created for performance analysis. Overall, the average runtime for adding a single structure to the database is 29 s (Figure 3a). This value does not increase through the different slices of data. That means that the process of adding a structure does not slow down with database size. The majority of the runtime (92%) is used for calculating the EDIA values. The highest runtimes in each data block are reached for structures with a very high number of atoms (≥60 000).

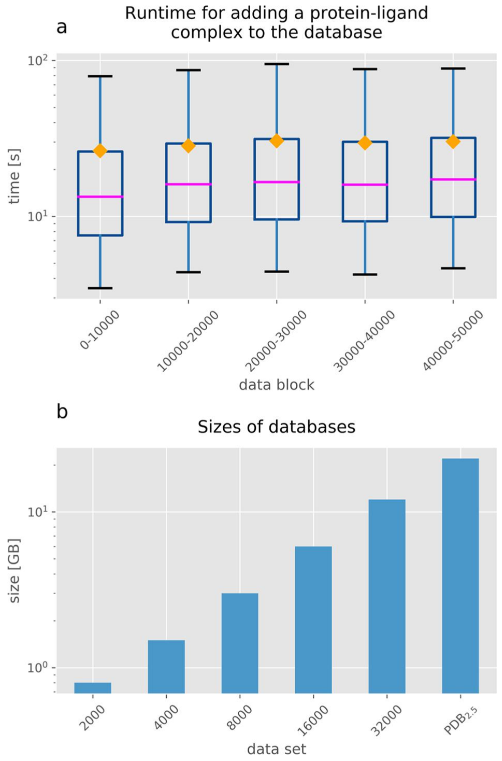


Figure 3. Performance measurements for database creation with *NAOMI*nova. (a) A whisker plot shows the time for adding one PDB file to the database for subsequent data slices of 10 000 structures. The blue box shows the lower and upper quartile, and the whiskers represent 90% of the results, from 5 to 95%. The median is shown as a magenta line. The mean is indicated by the yellow square. (b) The bar plot shows the disk size of databases with increasing numbers of PDB files.

In order to assess the dependency of the database size on the number of inserted PDB structures, sets with increasing numbers of randomly selected PDB structures were created from $PDB_{2.5}$. For each set, a *NAOMI*nova database was calculated (Figure 3b). The database size grows approximately linearly with the number of inserted PDB structures. Overall, a database containing all structures of $PDB_{2.5}$ has a size of 22 GB.

The performance for adding custom-defined substructures highly depends on the used SMARTS pattern and the chosen $EDIA_{min}$ value. Two main factors have to be considered concerning the runtime here: The more complex the SMARTS pattern, the longer the runtime of the matching algorithm. However, fewer results will be detected, and thus, fewer transformations have to be performed. As an example, three different substructures were registered and added to the database: (1) "CN1" with SMARTS pattern "C[NH2]", (2) "CN2" with SMARTS pattern "[C$(C[CR1])][NH2]", and (3) "CN3" with SMARTS pattern "[C$(CC1CCCCC1)]-[NH2]". All three substructures describe the same molecular structure but differ in their molecular context. For all three substructures, the 3D template molecule was defined using the SMILES C[NH2]. The $EDIA_{min}$ threshold was set to 0.8. The common molecular structure is schematically depicted in Figure 4a. The molecular contexts of the individual substructures are displayed in Figure 4b−d. In Figure 4e−g, the runtime needed for the data collection step is plotted against the number of hits for databases containing different numbers of PDB files.

As expected, more hits are detected for CN1 than for CN2 and CN3. For each of the different substructures, a linear dependency between the number of hits and the runtime can be observed. The linear regression curves are displayed as blue lines in Figure 4e−g. The slope of this curve increases from CN1 to CN2 and CN3, indicating that the detection of one hit is more time-consuming for a more complex SMARTS pattern than that for a simpler one.

The increase in disk size of a database by adding substructures only depends on the number of detected partner points. Roughly, one partner point adds 380 bytes to the database.

The runtime needed for querying partner points was assessed using the database containing the $PDB_{2.5}$ data set and the three substructures CN1, CN2, and CN3 from the previous step. For each of the added substructures, three different filtering steps were performed. First of all, all partner points for the respective substructure were queried. Second, only oxygens were requested. Eventually, the database was only filtered for oxygens derived from ligands. The runtime and the number of detected partner points are shown in Table 1.

As expected, the runtime for querying partner points from the database depends on the number of results. In general, queries with up to 1.9×10^7 results can be answered in less than 90 s. Most of the used queries can even be answered in less than 1 s. It can be concluded that *NAOMI*nova is able to support interactive analyses even if large data sets are used.

■ APPLICATION EXAMPLES

*NAOMI*nova can be applied in very heterogeneous scenarios. Its primary use is for the analysis of large, generic structure collections. Moreover, it can also be used for detailed analysis of smaller, focused collections summarizing the data available for a target class or even from a MD run on a single structure. Here, we will apply *NAOMI*nova on two data sets, the general set of the PDB-bind[21] and an ensemble of 408 carbonic anhydrase (CA) structures. For both scenarios, the used SMARTS expressions, the number of detected substructure hits, and the number of partner points are shown in Table 2. In both examples, the $EDIA_{min}$ of the substructures is set to 0.8,

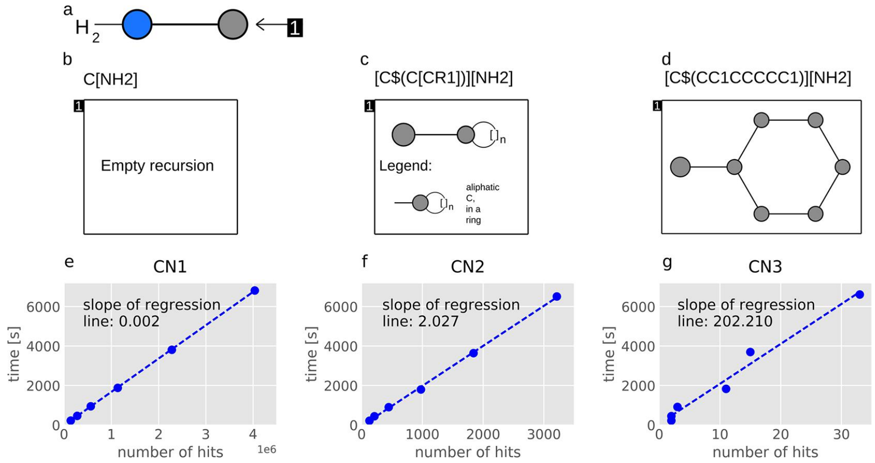


Figure 4. Substructure added to a *NAOMI*nova database for runtime measurements. (a) General depiction of the molecular structure represented by all three substructures (CNH_2). The carbon atom is indicated with "←1". This carbon atom is further specified by a recursive SMARTS shown in (b–d) in boxes. (b–d) Specific depiction of the recursion in the SMARTS pattern of CN1, CN2, and CN3, respectively. (b) The carbon is not further defined. (c) The carbon atom has to be connected to any aliphatic ring. (d) The carbon atom has to be connected to a cyclohexane. The large gray circle represents the carbon that is further defined by the different recursions. The SMARTS patterns were visualized using the SMARTSviewer.[20] (e–g) The number of hits is plotted against the overall runtime for adding CN1, CN2, and CN3 to the database, respectively. A linear regression curve is depicted as a blue dotted line in all three plots.

Table 1. Runtimes and Number of Received Partner Points (pps) for Querying Three Different Filters on Three Different Substructures Based on the $PDB_{2.5}$ Data Set along with Mean Values and Standard Deviations of the Run Times of Three Independent Experiments

filter criteria for partner points	CN1	CN2	CN3
no filter	1.9×10^7 pps,	12 930 pps,	140 pps,
	90 s ± 0.2 s	<1 s	<1 s
element type = oxygen	1.4×10^7 pps,	10 224 pps,	112 pps,
	67 s ± 0.1 s	<1 s	<1 s
element type = oxygen,	1.7×10^5 pps,	536 pps,	20 pps,
location = ligand	2.7 s ± 0.05 s	<1 s	<1 s

Table 2. Runtime and Database Size for Two Data Sets, PDB-bind and CA[a]

	PDB-bind (general set)	CA ensemble
number of structures in data set	13 308	408
DB size before entering substructures	4.7 GB	73 MB
DB size after entering substructures	22 GB	74 MB
number of substructure hits	carbonyl[b]: 5.5×10^6 planar carbonyl[b]: 5.5×10^6	Gln92[b]: 334 Thr199[b]: 335 Thr200[b]: 336
number of partner points	carbonyl[b]: 2.2×10^7 planar carbonyl[b]: 2.6×10^7	Gln92[b]: 3025 Thr199[b]: 2971 Thr200[b]: 1362
time for adding all substructures	1:23 h	1:03 min

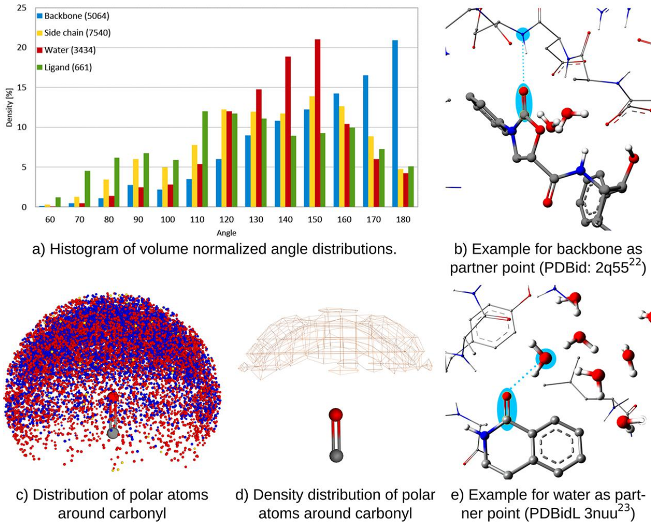
[a] Numbers are for structures with $EDIA_{min} \geq 0.8$; amino acid numbers correspond to PDBid 1zsb.[31] Further information about the substructure selection can be found in the section Application Examples. [b] SMARTS and SMILES pattern for carbonyl: O=C, 3D template: O=C; planar carbonyl: O=CC, 3D template: O=CC; Gln92: [NH2][C$(CCCC(NC(=O)CC(C)CC)C(=O)-NCCc1ccccc1)]=O, 3D template: [NH2]C=O; Thr199: [OH]C[C$(C(NC(=O)CCC(C)C)C(=O)NCC(C)O)], 3D template: [OH]CC; Thr200: OC[C$(C(NC(=O)CC(C)O)C(=O)-N1CCCC1)], 3D template: [OH]CC.

the suggested cutoff for atoms well supported by electron density.[8,9]

Analyzing Carbonyl Interaction Patterns. We use PDB-bind to demonstrate the potential of analyzing the interaction preference of a certain functional group across a large database, which can be linked to affinity data. As a concrete example, the overall distribution of atoms around a nonfurther-specified carbonyl group was employed. The question, *What are the interaction preferences of a carbonyl oxygen atom?* is addressed. Herein, the overall distribution of atoms around the carbonyl structure was analyzed. Using this example, we highlight the importance of carefully analyzing the geometry around functional groups. In this case, especially the significant difference between the distribution with and without the carbonyl plane is defined. To this end, a database with all protein–ligand structures from the general set of PDB-bind (version 2016) was compiled. The substructure "carbonyl" was defined with the SMARTS pattern "O=C" and an $EDIA_{min}$ of 0.8. The 3D template molecule was defined using the SMILES pattern O=C. In total, about 1.06×10^7 partner points were detected.

The distribution of all detected partner points in a hydrogen bond distance range (2.6–3.5 Å) to any atom of the substructure is shown in Figure 5c. Additional information about the density of the point distribution is visualized in Figure 5d. It can be seen that most of the partner points are above the oxygen atom even though single partner points can also be found further below. In this first analysis, the plane of the carbonyl is not defined. Therefore, the points are evenly distributed around the carbonyl group. However, more detailed

DOI: 10.1021/acs.jcim.7b00291



a) Histogram of volume normalized angle distributions.

b) Example for backbone as partner point (PDBid: 2q55[22])

c) Distribution of polar atoms around carbonyl

d) Density distribution of polar atoms around carbonyl

e) Example for water as partner point (PDBidL 3nuu[23])

Figure 5. Distribution of partner points around carbonyl functional groups in the PDB-bind. Filter criteria: ligand as the central structure location, distance to any atom 2.6−3.5 Å, only polar partner points (oxygen, nitrogen, sulfur); the substructure and corresponding partner point are highlighted in light blue.

analysis of the data reveals distinct differences among the different groups of partner points. Figure 5a shows the distribution of the angle between the carbonyl bond and the vector from the oxygen atom to a partner point separated for different groups of partner points. Herein, backbone atoms reveal a maximum in the 180° direction, that is, head-to-head to the oxygen atom and in between the two electron lone pairs (Figure 5b). The distribution of water molecules around carbonyl oxygen atoms on the other hand exhibits an optimum at 140−150°, which is in better agreement with the assumed electron lone pair directions at about 120° (Figure 5a). This difference in distribution might be due to the spatial restrictions of the partner points, that is, if water molecules have more space around the carbonyl functional group, they rather assemble in ideal interaction directions, whereas backbone atoms are more restricted in their position due to the overall protein structure (Figure 5b,e).

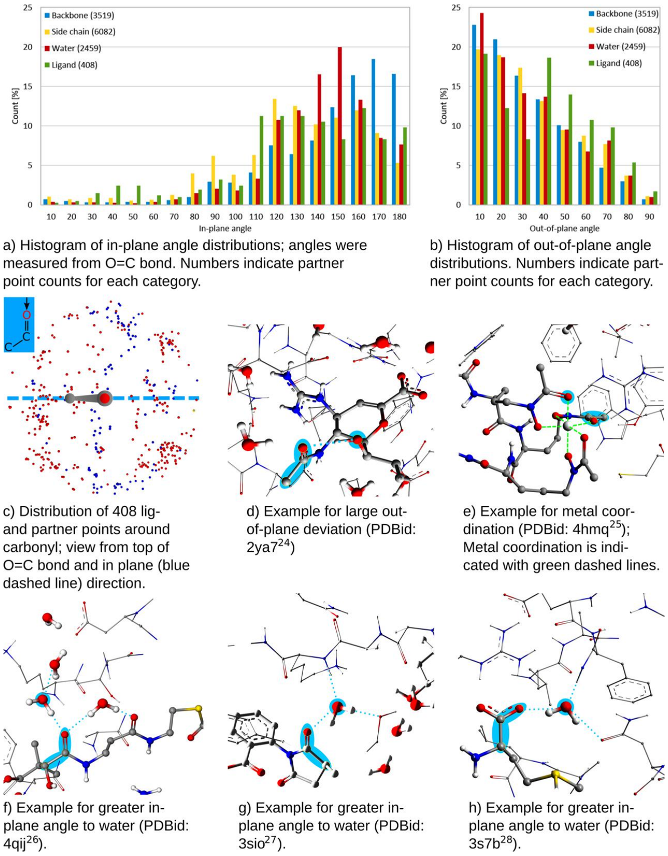
In a next step, the substructure was extended to a "planar carbonyl" using the SMARTS pattern "O═CC". Accordingly, the 3D template molecule was defined using the same expression. This way, a plane was defined and the partner point distribution with respect to in- and out-of-plane direction could be analyzed in greater detail. In total, about 2.6×10^7 partner points were detected for this substructure. This number is slightly higher than that for O═C because in the initial process of adding substructures to the database partner points are included when they are within 4.5 Å of any atom of the functional group.

Partner atoms from protein side chains and ligands have their preferred positions in approximate agreement with the lone pair directions. The in-plane angles show a plateau between 120 and 160°, while the out-of-plane angle has a clear maximum at 0° (Figure 6a,b). Noteworthy is the second peak at 40° in the out-of-plane distribution for ligands (Figure 6b,c). The underlying structures reveal both hydrogen bond interactions (Figure 6d) as well as close contacts due to metal coordination (Figure 6e). In accordance with the observation above, backbone atoms have their main peak for the in-plane angle between 160 and 180° (Figure 6a). Water molecules have their in-plane angle peak at 150° (Figure 6a). Closer examination reveals that most of those water molecules are integrated in an extended hydrogen bond network and form multiple hydrogen bonds (>two hydrogen bonds, Figure 6f−h). Therefore, they might be more restricted and form nonideal interactions trying to fulfill as many hydrogen bonds as possible.

The example of the carbonyl with and without plane definition shows the potential of *NAOMI*nova to geometrically analyze large amounts of data in great detail. The back-link function helps understand deviations from the expected, ideal distributions. Moreover, our example demonstrates the necessity of an interactive tool with back-link functionality. Protein structures as available today have a strong geometric bias. On the one hand, specific functional groups are highly overrepresented. On the other hand and even more important, functional groups appear in highly overrepresented structural motifs and protein folds. The back-link functionality is therefore indispensable for detecting artificial histogram peaks. A detailed analysis of hydrogen bond geometries for a large collection of functional groups has recently been performed with *NAOMI*nova.[19] One major aspect, which can also be observed in the shown carbonyl example, is the great deviation in out-of-plane direction. Bissantz et al.[29] specified the out-of-plane angle between the donor−hydrogen···acceptor as less than 30°. Because hydrogen atoms are not resolved in most crystal structures, we prefer to measure the out-of-plane angle between the donor, oxygen, and carbonyl carbon. Here, we found that the majority of donor···acceptor−carbon angles, which are hydrogen-atom-independent, are within 25°, with variations up to 80°. Similarly to the majority of our detected angles, Ippolito et al.[30] derived an out-of-place angle of 20−30°.

Carbonic Anhydrase Ensemble−Ligand Extension. CA catalyzes the reversible hydration of carbon dioxide to

DOI: 10.1021/acs.jcim.7b00291



a) Histogram of in-plane angle distributions; angles were measured from O=C bond. Numbers indicate partner point counts for each category.

b) Histogram of out-of-plane angle distributions. Numbers indicate partner point counts for each category.

c) Distribution of 408 ligand partner points around carbonyl; view from top of O=C bond and in plane (blue dashed line) direction.

d) Example for large out-of-plane deviation (PDBid: 2ya7[24])

e) Example for metal coordination (PDBid: 4hmq[25]); Metal coordination is indicated with green dashed lines.

f) Example for greater in-plane angle to water (PDBid: 4qij[26]).

g) Example for greater in-plane angle to water (PDBid: 3sio[27]).

h) Example for greater in-plane angle to water (PDBid: 3s7b[28]).

Figure 6. In- and out-of-plane angle distributions around carbonyl functional groups. Filter criteria: ligand as the central structure location, distance to carbonyl oxygen 2.6–3.5 Å, only polar partner points (oxygen, nitrogen, sulfur); the substructure and corresponding partner point are highlighted in light blue. 3D coordinates of hydrogen atoms were calculated with Protoss.[14]

bicarbonate.[32,33] Different CAs with different physiological roles are present in humans. Mis-regulations can cause diseases like glaucoma and cancer.

In order to answer the question *How could a fragment binding to CA be extended?*, we analyzed an ensemble of CA binding pockets and searched for amino acids frequently interacting with parts of known ligands. The ensemble data set of CA structures was taken from the NHSE set.[34] Briefly, it was generated using ASCONA[35] for structural alignment and SIENA[34] for binding site ensemble generation using 100% site identity and a resolution cutoff of 2.5 Å. Three common amino acids were identified that often directly interact with the ligand—two threonine side chains buried in the pocket and one glutamine side chain at the rim of the binding pocket (Thr199, Thr200, and Gln92 based on PDBid: 1zsb[31]). For each of these amino acids, a SMARTS pattern describing the interacting atoms (hydroxyl group or amide group) in the substructure part and the two neighboring amino acids in the molecular context was designed. For example, using the CC[OH] for mapping of the threonine residues leads to unambiguous mapping of the interacting oxygen atom. To distinguish between Thr199 and Thr200, the remaining amino acid and the neighboring ones (one left, one right) were defined as the molecular context using recursive SMARTS. This way, for each amino acid of interest, a tripeptide was defined that led to unambiguous matching in CA. Because the respective sequences of amino acids are unique in the set of CAs, only the targeted amino acids are hit during the substructure search. The exact SMARTS and SMILES patterns used to define the substructure and the 3D template molecules are given in Table 2.

Threonine 199 is conformationally restrained in the binding pocket (Figure 7a). Distinct patches for surrounding side chains—histidines (that coordinate the zinc) and asparagine—as well as a conserved water molecule can be observed (Figure 7g,j). Interestingly, threonine 199 has no backbone interactions at all within a 3.5 Å distance (Figure 7d). Almost always it interacts with a sulfonamide functional group of ligands, with little variations in its position (Figure 7m). Threonine 200 on the other hand has backbone atoms as surrounding partners but no side-chain atoms within a 3.5 Å distance (Figure 7e,h). Three tetrahedrally arranged patches of water molecules can be observed (Figure 7k), which are partially displaced by ligand atoms (Figure 7n). Glutamine 92 has distinct patches for the

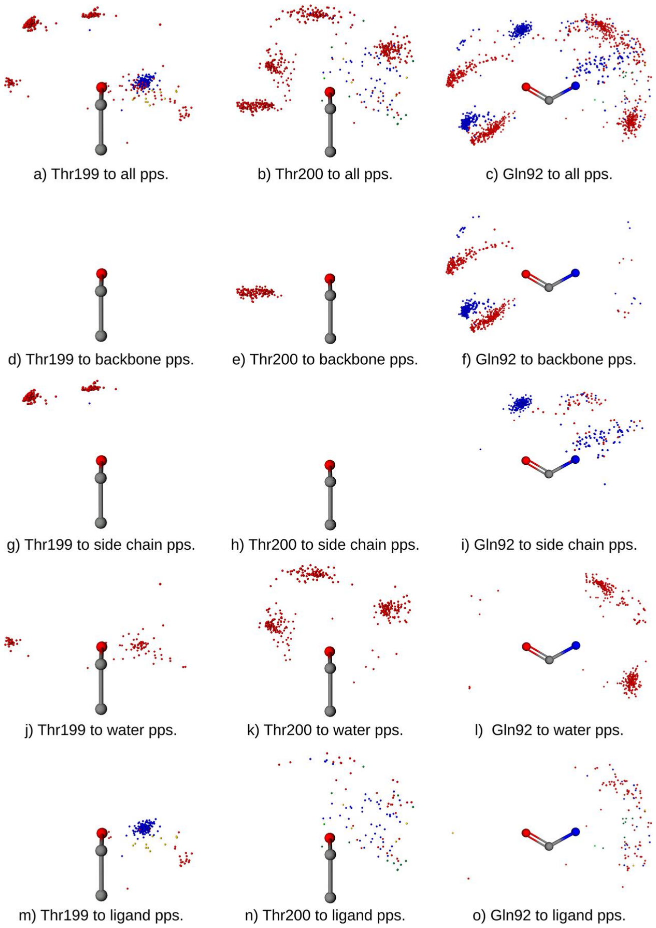


Figure 7. Distribution of partner points around three different amino acid side chains for CA. Amino acid numbers correspond to PDBid 1zsb; the CA database was filtered for atoms at 2.6−3.5 Å distance; pps = partner points.

surrounding backbone as well as side-chain atoms (Figure 7f,i). Herein, the oxygen atom of the primary amide always interacts with a histidine side chain. The nitrogen atom of the primary amide on the other hand interacts with water molecules (Figure 7l), which can be displaced by ligand atoms (Figure 7o).

In the uninhibited structure of CA, two water molecules coordinate the zinc in addition to the histidine side chains (Figure 8a). Those two water molecules are displaced by the sulfonamide group of the ligands (Figure 8b), which then coordinates the zinc and blocks the enzymatic function of CA.

We used *NAOMI*nova to analyze the functional groups that were used to displace the water molecules around the nitrogen of the primary amide of glutamine 92. Therefore, we superimposed the distribution of water and ligand atoms surrounding this glutamine into the CA pocket (Figure 8b). Using the back-link functionality, the respective structures that reach to the glutamine side chain were easily accessible (for example ligands, see Figure 8c−f). One of those structures revealed a dual-tail ligand, where a triazole group displaces the water molecule and a nitrogen acceptor of the triazole group interacts with the glutamine side chain (Figure 8f). Overall, different polar extensions have been used to displace the water molecule and interact directly with the glutamine side chain.

Here, *NAOMI*nova allows comprehensive analysis of the data, facilitates easy access to the corresponding structures, and gives ideas for potential extensions of the ligand in a drug development process. In principle, similar analyses could be done by simply superimposing structures of the same protein followed by visually inspecting the interactions with a ligand. However, with more than 100 structures, such a visual analysis can be difficult. In this respect, *NAOMI*nova provides the perfect means for spatial analyses of larger data collections.

■ CONCLUSIONS

In this publication, we presented a method for the interactive geometric analysis of preferred interaction directions of user-defined substructures in protein complexes. We integrated the method into a software tool named *NAOMI*nova. The great strength of our method is its flexibility concerning the used data sets, the analyzed substructures, and the filtering properties.

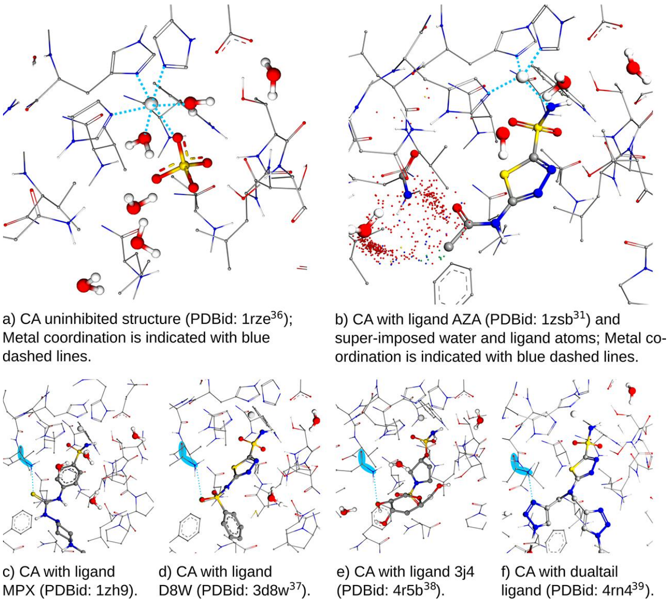

a) CA uninhibited structure (PDBid: 1rze[36]); Metal coordination is indicated with blue dashed lines.
b) CA with ligand AZA (PDBid: 1zsb[31]) and super-imposed water and ligand atoms; Metal co-ordination is indicated with blue dashed lines.
c) CA with ligand MPX (PDBid: 1zh9).
d) CA with ligand D8W (PDBid: 3d8w[37]).
e) CA with ligand 3j4 (PDBid: 4r5b[38]).
f) CA with dualtail ligand (PDBid: 4rn4[39]).

Figure 8. CA binding pocket: uninhibited and with different ligands displacing water molecules. The glutamine side chain is highlighted in light blue.

First, the set of analyzed PDB structures can be freely chosen. Even large data sets of up to 60 000 complexes can be handled. Second, the substructures can be defined using the powerful SMARTS language. Even though *NAOMI*nova was mainly developed with the purpose to analyze strong hydrogen bonds, also weak hydrogen bonds, that is, C−H···O/N, can be analyzed by defining a respective substructure containing the C−H of interest. Still, one remaining limitation is the analysis of interactions between carbons, such as $\pi-\pi$ interactions, beause carbon atoms are not stored as partner atoms. The pure number of carbons would cause extensive growth of the database, requiring substantial improvements of software and the used hardware. These enhancements are the subject of further development. Third, *NAOMI*nova has various filter options on atomic attributes, which can be used to tailor the distribution of partner points to specific needs. Herein, not only protein−ligand interactions can be analyzed in detail. *NAOMI*nova can also be used to analyze intraprotein interactions as well as protein−protein interactions. Another strength of *NAOMI*nova is its ability to assess the quality of the analyzed data, which allows one to increase the reliability of derived hypotheses. Here, EDIA values can be used to analyze only atoms well represented in the underlying experimental electron density. Moreover, structural anomalies can be easily traced back from every data point to its originating protein structure using the back-link functionality.

All calculated data are stored in one SQLite database, which does not require any server infrastructure and can be used on a standard desktop PC. Once calculated, the speed of the data retrieval system allows for interactive analysis of the data.

We presented two possible application scenarios showing how *NAOMI*nova can be used to analyze the spatial distribution of interaction partners around a functional group and how ideas can be deduced in order to chemically extend a fragment based on other ligands in the same binding site. Even more types of analyses are possible with *NAOMI*nova, including the distribution of spatial interactions in protein−protein interfaces or the analysis of interactions over time in MD simulations.

Overall, our method integrated in *NAOMI*nova enables structure-based data mining in the ever-growing Protein Data Bank. This way, knowledge about structural features is provided almost instantaneously with many applications in molecular design.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00291.

> List containing all available filter criteria, their possible values and their default values, a detailed introduction to the GUI, explanation of density grid calculation, histogram normalization, and calculation of EDIA$_m$ (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

ORCID

Eva Nittinger: 0000-0001-7231-7996
Kai Sommer: 0000-0003-1866-8247
Matthias Rarey: 0000-0002-9553-6531

Author Contributions

‡T.I. and E.N. contributed equally to this work.

Notes

The authors declare no competing financial interest.

All calculations shown here were performed on a single core on a standard PC equipped with an Intel I5-4570 (3.2 GHz) processor, 16 GB of main memory, and a Samsung 950 pro PCIe solid-state drive (512 GB, model nvme) with a btrfs file system and a standard configuration of a Linux openSUSE 13.1 distribution. *NAOMI*nova is part of the NAOMI ChemBio Suite and available for Linux and Windows from http://www.zbh.uni-hamburg.de/naominova, free for academic use and evaluation purposes. All feedback is greatly appreciated and supports the further development of *NAOMI*nova. Structural preprocessing functionality including Protoss and EDIA are also available as a web service at http://proteins.plus.

■ ACKNOWLEDGMENTS

The author thank Gudrun Lange and Robert Klein from Bayer CropScience for fruitful discussions during the analysis of interaction geometries.

■ REFERENCES

(1) Panigrahi, S. K.; Desiraju, G. R. *Proteins: Struct., Funct., Genet.* **2007**, *67*, 128−141.

(2) Zimmermann, M. O.; Lange, A.; Zahn, S.; Exner, T. E.; Boeckler, F. M. *J. Chem. Inf. Model.* **2016**, *56*, 1373−1383. PMID: 27380316.

(3) 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* **2000**, *28*, 235−242.

(4) Bruno, I. J.; Cole, J. C.; Lommerse, J. P.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525−537.

(5) Allen, F. H. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380−388.

(6) *IsoStar User Guide and Tutorials*, Cambridge Crystallographic Data Centre, 2017 CSD Release. http://isostar.ccdc.cam.ac.uk/docs/isostar/isostar.html (accessed March 2017).

(7) Verdonk, M. L.; Cole, J. C.; Watson, P.; Gillet, V.; Willett, P. *J. Mol. Biol.* **2001**, *307*, 841−859.

(8) Nittinger, E.; Schneider, N.; Lange, G.; Rarey, M. *J. Chem. Inf. Model.* **2015**, *55*, 771−83.

(9) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. 2017, unpublished results.

(10) Urbaczek, S.; Kolodzik, A.; Groth, I.; Heuser, S.; Rarey, M. *J. Chem. Inf. Model.* **2013**, *53*, 76−87. PMID: 23176552.

(11) Schomburg, K. T.; Bietz, S.; Briem, H.; Henzler, A. M.; Urbaczek, S.; Rarey, M. *J. Chem. Inf. Model.* **2014**, *54*, 1676−1686. PMID: 24851945.

(12) Hilbig, M.; Rarey, M. *J. Chem. Inf. Model.* **2015**, *55*, 2071−2078. PMID: 26389652.

(13) Inhester, T.; Bietz, S.; Hilbig, M.; Schmidt, R.; Rarey, M. *J. Chem. Inf. Model.* **2017**, *57*, 148−158. PMID: 28128948.

(14) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. *J. Cheminf.* **2014**, *6*, 12.

(15) Daylight SMARTS examples, Daylight Chemical Information Systems, Inc.. http://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html (accessed March 2017).

(16) Weininger, D. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(17) Sommer, K.; Friedrich, N.-O.; Bietz, S.; Hilbig, M.; Inhester, T.; Rarey, M. *J. Chem. Inf. Model.* **2016**, *56*, 1105−1111. PMID: 27227368.

(18) Ehrlich, H.-C.; Rarey, M. *J. Cheminf.* **2012**, *4*, 13.

(19) Nittinger, E.; Inhester, T.; Bietz, S.; Meyder, A.; Schomburg, K. T.; Lange, G.; Klein, R.; Rarey, M. *J. Med. Chem.* **2017**, *60*, 4245−4257. PMID: 28497966.

(20) Schomburg, K.; Ehrlich, H.-C.; Stierand, K.; Rarey, M. *J. Chem. Inf. Model.* **2010**, *50*, 1529−1535. PMID: 20795706.

(21) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. *Bioinformatics* **2015**, *31*, 405−12.

(22) Reddy, G. S. K. K.; Ali, A.; Nalam, M. N. L.; Anjum, S. G.; Cao, H.; Nathans, R. S.; Schiffer, C. A.; Rana, T. M. *J. Med. Chem.* **2007**, *50*, 4316−4328. PMID: 17696512.

(23) Medina, J. R.; Blackledge, C. W.; Heerding, D. A.; Campobasso, N.; Ward, P.; Briand, J.; Wright, L.; Axten, J. M. *ACS Med. Chem. Lett.* **2010**, *1*, 439−442.

(24) Gut, H.; Xu, G.; Taylor, G. L.; Walsh, M. A. *J. Mol. Biol.* **2011**, *409*, 496−503.

(25) Cheng, W.; Li, Q.; Jiang, Y.-L.; Zhou, C.-Z.; Chen, Y. *PLoS One* **2013**, *8*, e71451.

(26) Song, H.; Sung, H. P.; Tse, Y. S.; Jiang, M.; Guo, Z. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2014**, *70*, 2959−2969.

(27) Nemecz, k.; Taylor, P. *J. Biol. Chem.* **2011**, *286*, 42555−42565.

(28) Ferguson, A.; Larsen, N.; Howard, T.; Pollard, H.; Green, I.; Grande, C.; Cheung, T.; Garcia-Arenas, R.; Cowen, S.; Wu, J.; Godin, R.; Chen, H.; Keen, N. *Structure* **2011**, *19*, 1262−1273.

(29) Bissantz, C.; Kuhn, B.; Stahl, M. *J. Med. Chem.* **2010**, *53*, 5061−5084.

(30) Ippolito, J. A.; Alexander, R. S.; Christianson, D. W. *J. Mol. Biol.* **1990**, *215*, 457−471.

(31) Huang, C.-c.; Lesburg, C. A.; Kiefer, L. L.; Fierke, C. A.; Christianson, D. W. *Biochemistry* **1996**, *35*, 3439−3446. PMID: 8639494.

(32) Krishnamurthy, V. M.; Kaufman, G. K.; Urbach, A. R.; Gitlin, I.; Gudiksen, K. L.; Weibel, D. B.; Whitesides, G. M. *Chem. Rev.* **2008**, *108*, 946−1051. PMID: 18335973.

(33) Pinard, M. A.; Mahon, B.; McKenna, R. Probing the Surface of Human Carbonic Anhydrase for Clues towards the Design of Isoform Specific Inhibitors. *BioMed Res. Int.* **2015**, *2015*, 1.

(34) Bietz, S.; Rarey, M. *J. Chem. Inf. Model.* **2016**, *56*, 248−259.

(35) Bietz, S.; Rarey, M. *J. Chem. Inf. Model.* **2015**, *55*, 1747−1756.

(36) Håkansson, K.; Wehnert, A.; Liljas, A. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *50*, 93−100.

(37) Genis, C.; Sippel, K. H.; Case, N.; Cao, W.; Avvaru, B. S.; Tartaglia, L. J.; Govindasamy, L.; Tu, C.; Agbandje-McKenna, M.; Silverman, D. N.; Rosser, C. J.; McKenna, R. *Biochemistry* **2009**, *48*, 1322−1331. PMID: 19170619.

(38) Moeker, J.; Mahon, B. P.; Bornaghi, L. F.; Vullo, D.; Supuran, C. T.; McKenna, R.; Poulsen, S.-A. *J. Med. Chem.* **2014**, *57*, 8635−8645. PMID: 25254302.

(39) Tanpure, R. P.; Ren, B.; Peat, T. S.; Bornaghi, L. F.; Vullo, D.; Supuran, C. T.; Poulsen, S.-A. *J. Med. Chem.* **2015**, *58*, 1494−1501. PMID: 25581127.

Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein-Ligand Interfaces.

[D5] **Nittinger, E.**; Inhester, T.; Bietz, S.; Meyder, A.; Schomburg, K.; Lange, G.; Klein, R.; Rarey, M. Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein-Ligand Interfaces. J. Med. Chem. 2017, 60 (10): 4245-4257.

http://pubs.acs.org/articlesonrequest/AOR-K746IyzEeHqWswwz7eCC

Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein–Ligand Interfaces

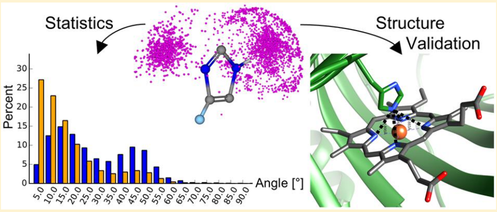
Eva Nittinger,[†,∥] Therese Inhester,[†,∥] Stefan Bietz,[†] Agnes Meyder,[†] Karen T. Schomburg,[†,§] Gudrun Lange,[‡] Robert Klein,[‡] and Matthias Rarey*[,†]

[†]Universität Hamburg, ZBH—Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg, Germany

[‡]Bayer CropScience AG, Industriepark Hoechst, G836, 65926 Frankfurt am Main, Germany

Supporting Information

ABSTRACT: Protein–ligand interactions are the fundamental basis for molecular design in pharmaceutical research, biocatalysis, and agrochemical development. Especially hydrogen bonds are known to have special geometric requirements and therefore deserve a detailed analysis. In modeling approaches a more general description of hydrogen bond geometries, using distance and directionality, is applied. A first study of their geometries was performed based on 15 protein structures in 1982. Currently there are about 95 000 protein–ligand structures available in the PDB, providing a solid foundation for a new large-scale statistical analysis. Here, we report a comprehensive investigation of geometric and functional properties of hydrogen bonds. Out of 22 defined functional groups, eight are fully in accordance with theoretical predictions while 14 show variations from expected values. On the basis of these results, we derived interaction geometries to improve current computational models. It is expected that these observations will be useful in designing new chemical structures for biological applications.



■ INTRODUCTION

Hydrogen bonds (H-bonds) are fundamental interactions in protein–ligand complexes and are the main reason for protein–ligand selectivity. A thorough understanding of H-bond geometries is mandatory in medicinal chemistry for structure-based design approaches. Many modeling applications related to protein–ligand complexes, ranging from docking and scoring to detailed analysis of binding events and catalytic function, are based on H-bonds. Therefore, computational tools relying on interactions and theoretical considerations need an accurate geometric representation of H-bond interactions in order to retrieve relevant results for drug and pesticide development.

The conceptual model of an H-bond was first developed in the 1920s. Since then, intensive research has been conducted, ranging from structural studies on α-helices[1] and β-sheets[2] in proteins to the determination of H-bond geometries[3−21] and to studies exploring intramolecular hydrogen bonding in ligand design.[22,23]

Theoretical models for H-bond geometries have been derived from molecular structure theories. The valence shell electron pair repulsion (VSEPR)[24,25] model is one of the best accepted models, and all our results described herein were compared to this model. The VSEPR model assumes that the valence electron pairs surrounding an atom will repel each other and adopt a geometrical arrangement that minimizes this repulsion.

Two main questions are addressed in the present study. First, what are the preferred H-bond geometries that are observed in protein–ligand structures? Second, are the preferred geometries derived from experimental structures in accordance with the theoretical predictions based upon the VSEPR model?

Frequency distributions of H-bond distances and angles have been calculated from various data sets of experimental structures: small molecule crystal structure data from the Cambridge Structural Database (CSD),[26] neutron diffraction data, which include the hydrogen atom positions,[8] or protein data from the PDB.[27] They focused on the analysis of H-bonds formed by backbone amides,[3,9,19] distinguishing backbone from side chain H-bonds,[16,18] the description of oxygen H-bonds[4,5,7,11] or on other functional groups[8,10,12,15] (see Table 1). Since the last study using data from the PDB,[21] the amount of data in the PDB has increased dramatically and enables a substantiated new analysis.

Since protein–ligand complexes and their interactions are our main focus, we used the large number of high-resolution protein structures currently available in the PDB. We defined 22 typical H-bond acceptor and donor functions for the generation of comprehensive geometric interaction distributions. Since the hydrogen position is rarely available from the X-ray crystallography experiment, we have not taken the hydrogen position into account for the determination of interacting angles.

The selected 22 functional groups were analyzed with respect to their distance from surrounding polar atoms, i.e., oxygen or

Received: January 24, 2017
Published: May 12, 2017

DOI: 10.1021/acs.jmedchem.7b00101
J. Med. Chem. 2017, 60, 4245−4257

Table 1. Overview of Previous Studies on H-Bond Geometries[a]

year	author	data (source)	differentiation of functional groups	additional information and conclusions
1975	Kroon et al.[4]	45 (mols)	N	Analysis of 196 O−H···O H-bonds Majority of H-bonds is not linear
1979	Vinogradov[6]	95 (PDB)	Y	Analysis of 439 H-bonds Linear H-bonds X−H···Y > 150°
1981	Ceccarelli et al.[7]	24 (ND)	Y	Analysis of 100 O−H···O H-bonds 25% of H-bonds are bifurcated
1982	Jeffrey et al.[8]	32 (ND)	Y	Analysis of 168 H-bonds Charged groups are more often bifurcated than uncharged ones
1984	Taylor et al.[9]	889 (CSD)	Y	Analysis of 1509 N−H···O=C H-bonds Short H-bonds are more linear
1984	Baker et al.[10]	15 (PDB)	Y	Differentiation of SC Almost all donor functions are fulfilled
1984	Murray-Rust et al.[11]	NA (CCDF)	Y	H-bond geometries to oxygen Majority of partners in electron lone pair direction
1990	Ippolito et al.[12]	50 (PDB)	Y	Differentiation of SC Planar groups tend to form H-bonds 2−30° out of plane
1991	Preissner et al.[13]	13 (PDB)	N	Distinction of SS elements (α-helices, β-sheets) 25% of all H-bonds bifurcated (higher in SS)
1992	Sticke et al.[14]	42 (PDB)	Y	Derive conclusions for protein folding H-bond angles coincide with atom hybridization On average 1.1 H-bonds per residue
1996	Mills et al.[15]	48 000 (CSD)	Y	Differentiation of 39 FG Shorter H-bonds have a higher directionality
2003	Kortemme et al.[16]	698 (PDB)	N	Derivation of H-bond potentials BB, SC separation Calculation of hydrogen atoms with CHARMM19
2004	Molcanov et al.[17]	230 (CSD)	Y	Analysis of 230 ester interactions H-bond geometry depends on syn/anti arrangement
2004	Sarkhel et al.[18]	28 (PDB)	N	BB, SC separation Bifurcated H-bonds are longer Strong interactions in proteins show deviations from a linear geometry
2007	Podtelezhnikov et al.[19]	247 (PDB)	N	Machine learning approach to derive H-bond potentials Analysis in interpeptide BB H-bonds Optimization of H-bond geometry and strength
2007	Panigrahi et al.[20]	251 (PDB)	N	Analysis of NH···O, O−H···O, C−H···O H-bonds in PL complexes Calculation of hydrogen atoms with MOE and MMFF94x
2008	Liu et al.[21]	4535 (PDB)	Y	Derivation of distance- and angle-dependent H-bond potentials by potential of mean force approach Placement of hydrogen atoms

[a] Specification of data source: mols = small molecules, PDB = protein structures from the PDB, ND = neutron diffraction data, CSD = small molecules from the CSD, CCDF = Cambridge structural data file; FG = functional group, SS = secondary structure, SC = side chain, BB = backbone, PL = protein−ligand.

nitrogen atoms.[28] Furthermore, the angular distribution of polar atoms surrounding the functional groups has been analyzed extensively. Both consistencies and differences to the VSEPR model were observed. In this paper, we discuss the retrieved H-bond angles in protein−ligand complexes, including a detailed analysis of single cases diverging from the interaction geometries predicted by the VSEPR model. Our studies identified new geometric features important for the formation of H-bonds between protein and ligand. These and the special case analyses have enabled us to construct an improved computational model for hydrogen bonding interactions, which is expected to be of value to medicinal chemists.

■ MATERIALS AND METHODS

We extracted two different data sets, one from high and one from medium resolution structures within the PDB. The analysis focused on functional groups relevant to protein structures and was extended to functional groups often occurring in ligands.

Data Sets. Data quality obviously plays a crucial role in the analysis of molecular interactions on the atomic level. Additionally, the number of structures should be high in order to retrieve statistically reliable results. We therefore extracted protein−ligand structures from the PDB using the following advanced search criteria: experimental method = X-ray, molecule type = protein (exclusion of RNA and DNA), chain length ≥ 50 amino acids, and ligand = true (search date, April 27, 2016). HR set: a high-resolution protein−ligand set with resolution of ≤1.5 Å, resulting in 8783 structures. MR set: a medium-resolution protein−ligand subset with resolution of ≤2.5 Å, resulting in 65 266 structures.

The HR set forms the basis for the characterization of functional groups frequently occurring in protein−ligand

DOI: 10.1021/acs.jmedchem.7b00101

structures. The MR set forms an additional data pool in case the number of interaction partners for a functional group taking the HR set alone was too low for statistical analysis (less than 350 data points for each hydrogen atom or electron lone pair).

Selection of Functional Groups. Table 2 summarizes all functional groups considered in this analysis. Details about the detection of functional groups in protein structures can be found in the Supporting Information section S1. For each functional group and its binding partner the structural unit to which it belongs and which is either protein backbone (BB-unit), protein side chain (SC-unit), or ligand (Lig-unit) was kept for analysis. In this way, possible statistical bias due to structural artifacts can be tracked and considered upon analysis. In addition, the units to which the interacting partners belong are listed, which is either ”Lig-unit” or “Any-unit”.

Table 2. Functional Groups Used for Evaluation[a]

functional group	can be Don	can be Acc	unit	partner unit
amide, primary	y	y	Lig	Any
			SC	Lig
amide, secondary (*E* or *Z*)	y	y	Lig	Any
			BB	Lig
amide, tertiary	n	y	Lig	Any
amine, primary	y	y	Lig	Any
amine, secondary	y	y	Lig	Any
nitrogen (sp^2), primary	y	n	Lig	Any
nitrogen (sp^2), secondary	y	n	Lig	Any
amine, tertiary	n	y	Lig	Any
carbonyl	n	y	Lig	Any
carboxyl	n	y	Lig	Any
			SC	Lig
ether	n	y	Lig	Any
ester	n	y	Lig	Any
guanidine	y	n	Lig	Any
	y	n	SC	Lig
hydroxyl	y	y	Lig	Any
			SC	Any
enole	y	y	Lig	Any
			SC	Lig
imidazole	y	y	SC	Lig
imine	y	y	Lig	Any
nitrile	n	y	Lig	Any
phosphate	y	y	Lig	Any
sulfate	y	y	Lig	Any
sulfonamide	y	y	Lig	Any

[a] The H-bond donor and acceptor functionality can depend on the protonation state of the functional group (can be Don, can be Acc); unit = structural unit, the functional group belongs to: Partner unit = structural unit, the interacting atoms surrounding the functional group belong to Lig = ligand; SC = side chain; BB = backbone; Any = any surrounding nitrogen or oxygen atom belonging to: ligand, side chain, backbone, or water.

Structural Alignment of Functional Groups. In order to enable a comparison of the H-bond geometries, all detected functional groups were superimposed on a template structure of the respective functional group. The alignment allows the rejection of ambiguous geometries, the equal processing of functional groups for further analysis, and an easy visualization. The coordinates of the atoms within the templates have been chosen using ideal geometries according to the VSEPR model calculated within the NAOMI framework.[29] All template structures can be found in the Supporting Information section S2.

As a first step, an atom-mapping to the corresponding template was calculated. Since the exact location of delocalized bonds and hydrogen positions cannot be reliably conducted from most PDB files, atoms were considered as equal if their element type and their number of connected heavy atoms were identical. Hydrogen atoms were not used during this mapping step. A more precise explanation of the mapping procedure for functional groups can be found in the Supporting Information section S3 and Figure S1.

An rmsd-optimal superposition was determined using the Kabsch algorithm.[30] If the best superposition had an rmsd ≤ 0.2 Å, all nitrogen and oxygen atoms are selected that are closer than 4.5 Å to any atom of the functional group and are at least five bonds apart from any polar atom of the functional group. The structural unit (ligand, side chain, or backbone) was stored for each functional group in a SQLite database. Additionally, information about the partner atoms was stored: element and valence state in order to determine whether an atom might be an acceptor or donor, functional group (if possible), residue type (if possible), coordinate, structural unit (ligand, side chain, backbone, or water), minimal distance to any atom of the functional group, and covalent relation to the functional group, i.e., intra- or intermolecular.

All angles calculated in the course of this study were derived using the template coordinates of the functional groups and the transformed coordinates of partner atoms stored in the database.

Hydrogen Bond Angles. Atom pairs were analyzed in two different ways. A single angle was measured relative to the direction of the neighboring heavy atoms (see Figure 1). In the special case of single-bonded sp^2 hybridized oxygen atoms, in- and out-of-plane angles were determined (see Figure 2).

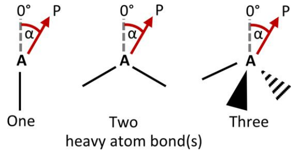


Figure 1. Measured angles for atom pairs depend on the number of heavy atom bonds: A = polar atom; P = polar partner atom. Angle between heavy atom and 0° direction: one heavy atom (HA) = 180°, two HA = 120°, three HA = 109.5°.

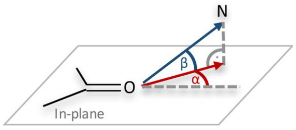


Figure 2. In-plane (red) and out-of-plane (blue) measurement for sp^2 hybridized oxygen atoms.

Frequency Distributions. Frequency distributions of H-bond distances and angles are represented by histograms. Histograms were generated for each polar atom of a functional group. Partner atoms are only considered if an H-bond could be actually formed; i.e., the histogram for the hydroxyl donor is generated only with potential acceptor partner positions and the other way around. The decision as to whether an atom can be a donor or an acceptor is based on the NAOMI atom types, an extension of the valence state.[29,31] This way, aromaticity as well as existing resonance forms can be described. Using these atom types, one can derive the number of lone pairs as well as the number of potential hydrogen atoms. For some functional groups this can lead to ambiguities. For instance, the nitrogen atoms of imidazoles can be either a donor or an acceptor depending on the tautomeric form. In these cases, the center group was treated once as donor and once as acceptor and both histograms were analyzed.

The frequency distributions were determined by counting the experimental structures with values falling into equidistant intervals of H-bond distances or angles. In order to obtain a measure for the relative probabilities that allow us to identify the preferred distances and angles, we have to divide these counts by the volume spanned by a particular interval. The resulting quantity has the dimension count per volume, and we therefore refer to it as density. The volume (V_i) of a distance interval is calculated using formula 1, that of an angle interval using formula 2.

$$V_i = \frac{4}{3}\pi({r_i}^3 - {r_{i-1}}^3) \tag{1}$$

$$V_i = \frac{2}{3}\pi|\cos(\alpha_{i-1}) - \cos(\alpha_i)|R \tag{2}$$

$$D_\mathrm{i} = \frac{\mathrm{count}_i}{V_i} \tag{3}$$

with $[r_{i-1}, r_i]$ being the ith distance interval in Å, $[\alpha_{i-1}, \alpha_i]$ the ith angle interval in radian, and R being an arbitrary cutoff distance. For the sake of comparability, counts and densities in the histograms are given in percent.

Nomenclature. Throughout this section the following nomenclature will be used (see Table 3):

Unit: the structural unit of the functional group.

Partner unit: the structural unit of potentially interacting atoms surrounding the functional group.

Example: primary amide N-SC-unit to Lig-partner unit describes a nitrogen atom of a primary amide functional group derived from an amino acid side chain surrounded by nitrogen or oxygen atoms from ligands.

RESULTS AND DISCUSSION

Occurrence of Functional Groups. A minimum of 95% of interacting pairs could successfully be superimposed on the template for each functional group (see Table S1). In order to draw meaningful conclusions, at least 350 interacting partner atoms for each donor or acceptor function of a functional group had to be available for the analysis. If the HR set provided less than 350 pairs, the MR set was included. This was mostly the case for groups that occur in ligands only. If still less than 350 data points were available, the functional group was excluded from further evaluation. Detailed results can be found in Table S2.

Table 3. Nomenclature for the Results Section[a]

functional group			surrounding	
atom	structural unit		partner unit	
	-Lig			
O			Lig	
	-SC	-unit		-partner unit
N			Any	
	-BB			

[a] Atom specification is used if the functional group contains oxygen and nitrogen atoms. Lig = ligand, SC = side chain, BB = backbone, Any = any surrounding nitrogen or oxygen atom belonging to ligand, side chain, backbone, or water.

Coverage of Hydrogen Bonds. The MR data set was used to calculate the percentage of hydrogen bonds covered using the 22 defined functional groups. Herein, every oxygen and nitrogen atom with either an electron lone pair or an attached hydrogen were classified as acceptor or donor, respectively.[29,31] Using our 22 defined functional groups, 99% of all nitrogen and oxygen donors and acceptors were covered (see Table 4). A separate analysis of nitrogen and oxygen atoms reveals that oxygen atoms are well covered in protein and ligand structures. However, nitrogen atoms show a lower percentage, which is for both protein and ligand due to cyclic nitrogen atoms, especially to those in aromatic rings. Ring systems have a great influence on the hydrogen bond functionality. Therefore, we deliberately chose histidine as a representative for aromatic nitrogen atoms. The nitrogen atoms of the histidine ring can be acceptor or donor and are abundant in protein structures. The analysis of aromatic nitrogen atoms in aromatic rings further supports the selection of histidine as a representative. In both cases (five- and six-membered aromatic rings) the interaction geometries agree very well with the observed geometries from histidine (Figure S3, Figure 9, and Figure 10).

Distance Distribution. The distances between a polar atom of a structural unit and its partner atoms were analyzed to define the cutoff distance between two polar atoms. Within this distance we can assume that neighboring atom pairs are indeed H-bonded. Outside this distance we can exclude pairs of atoms, which are close neighbors without forming a hydrogen bond.

The maximum in the distributions of different polar partner atoms around functional groups shows only small variations between 2.7 and 3.0 Å (Figure 3a, for all distance histograms see Figure S2). The distance distribution of carbon atoms was used to define a cutoff for the maximal H-bond distance considered here (see Figure 3b). The number of carbon atoms increases with distances greater than 3.0 Å. In order to exclude potentially non-hydrogen bonding atoms, we restricted therefore the atom−atom distance for the analysis of angular distributions between 2.6 and 2.9 Å.

Angle Distribution. Quantum mechanical calculations have shown that hydrogen-bond energies show a strong directionality.[32] Therefore, the analysis of angular distributions is of special interest. Out of the 22 functional groups, eight are fully in accordance with the theoretical considerations from the VSEPR model, i.e., hydroxyl, imidazole, guanidinium, and all amines (see Figure 4 for an example). Overall, donor directions agree better with the VSEPR model than acceptor directions. This section, however, will focus on those functional groups deviating from previous studies or theoretical considerations. Detailed information for all functional groups can be found in the Supporting Information.

Angle Distributions at sp^2 Oxygen Atoms. According to the VSEPR model, the expected electron lone pair directions of sp^2 oxygen atoms should exhibit an in-plane angle of about 60° and out-of-plane angles of 0° to accommodate the trigonal planar geometry (see Figure 5).

Table 4. Percentage of H-Bonds Covered with the 22 Functional Groups Used in This Study Based on the MR Data Set with overall 101 522 836 Atoms with Donor or Acceptor Function[a]

H-bond		location (acyclic)			location (cyclic)		
atom	type	protein	ligand	total	protein	ligand	total
N	Acc	9.5 (10.7)	2.1 (5.7)	5.2 (7.8)	32.4 (89.3)	5.6 (94.3)	16.8 (92.2)
	Don	91.6 (91.7)	40.6 (48.7)	91.5 (91.5)	7.1 (8.3)	11.3 (51.3)	7.2 (8.5)
O	Acc	99.9 (99.9)	92.2 (95.1)	99.7 (99.8)	0.1 (0.1)	4.7 (4.9)	0.2 (0.2)
	Don	98.9 (98.9)	88.9 (90.2)	98.0 (98.2)	1.1 (1.1)	9.7 (9.8)	1.8 (1.8)
total	Acc	99.7 (99.7)	83.4 (86.4)	99.4 (99.4)	0.2 (0.3)	4.8 (13.6)	0.3 (0.6)
	Don	92.4 (92.4)	76.7 (79.7)	92.2 (92.2)	6.5 (7.6)	10.1 (20.3)	6.5 (7.8)
	All	96.0 (96.1)	80.9 (83.9)	95.8 (95.8)	3.4 (3.9)	6.8 (16.1)	3.4 (4.2)

[a]Numbers indicate the percentage of the interacting atoms that are covered by the used 22 functional groups. Numbers in parentheses give the percentage from the total number of interacting atoms (101 522 836) in each category.

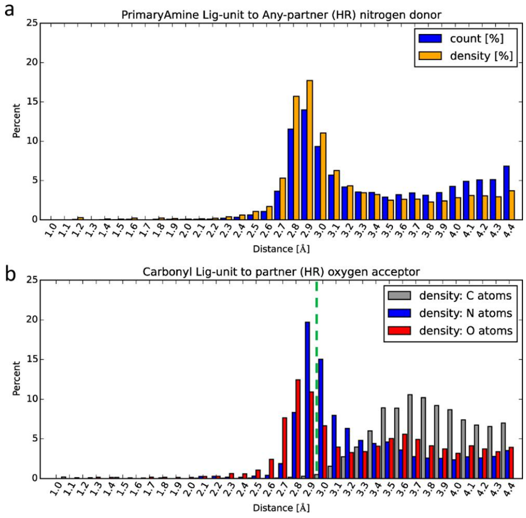


Figure 3. Distance histogram of (a) a primary amide Lig-unit to a Any-partner unit and (b) a carbonyl Lig-unit to Any-partner unit. Green dashed line indicates upper cutoff value for atom−atom distances.

The VSEPR geometry is best matched by the carboxyl group, which showed a mean in-plane angle of about 55°. The out-of-plane angle shows deviations of up to 25° (see Figure 5b and Figure 5c). All other angle distributions of sp^2 hybridized oxygens show greater variations from the ideal lone pair directions.

The angular distributions of some functional groups have well-defined maxima, e.g., secondary amides from the ligand unit (see Figure 5g and Figure S4), while the maxima in the distribution around esters (see Figure 5d) and tertiary amides (see Figure 5i) are rather diffuse. This is partly due to the number of interaction points available for the analysis. Noticeable are the similar angle distributions of ketones and primary amides (see Figure 5a,e,f). These functionalities deviate on average by 10−15° in their in-plane angle from the geometry expected by the VSEPR model. In addition, the angles of the sp^2 oxygens in secondary and tertiary amides and esters also group together (see Figure 5d,g−i). These three functional groups deviate significantly from the ideal in-plane lone pair direction with a maximum in the angular distribution between 30° and 35°. Additionally, they have a significant number of partner atoms located between the two electron lone pairs, i.e., at an in-plane angle close to 0°. A potential explanation for the deviations might be bifurcated hydrogen bonds. However, since hydrogen atoms are mostly unavailable in protein crystal structures, no definite answer can be given. Furthermore, the double bond “pushes” the electron lone pairs slightly together,[33] which might be a reason for the maximum at 55° for the carboxylate. This effect might cause slight differences from the ideal interaction direction but will certainly not lead to deviations up to 30°. High resolution structures with outliers have been examined, and example structures can be found in Figure S5.

Another observation is a greater deviation of the out-of-plane angle compared to the in-plane angle for all analyzed groups. Due to symmetry, the out-of-plane would need to be mirrored on the *x*-axis of Figure 5 to display its whole range (see Figure 6). Previous studies[7,12] observed that more partner points are

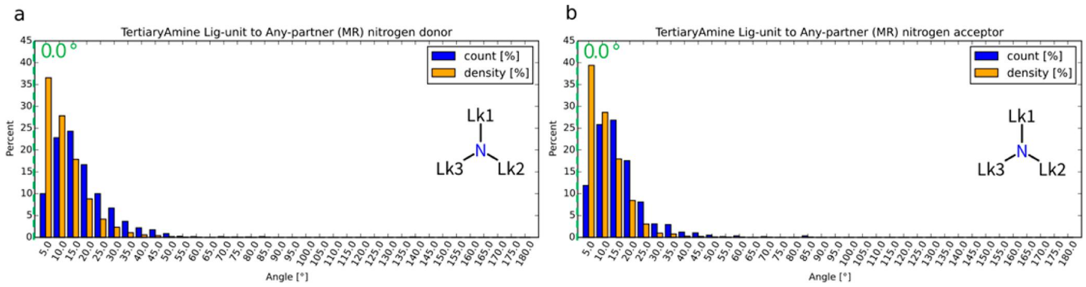


Figure 4. Histograms of tertiary amine functional groups: (a) of the nitrogen donor; (b) of the nitrogen acceptor.

DOI: 10.1021/acs.jmedchem.7b00101
J. Med. Chem. 2017, 60, 4245−4257

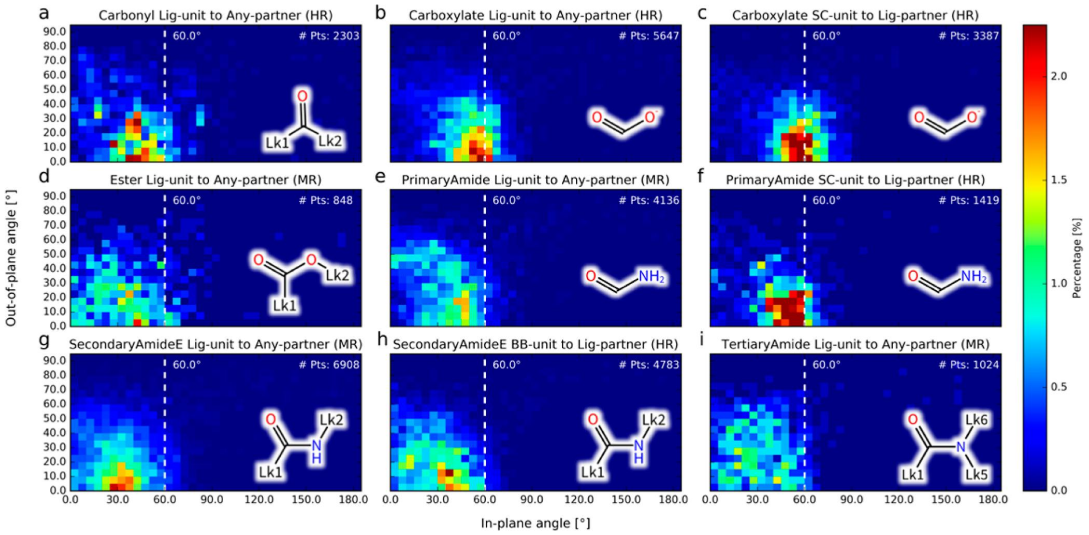


Figure 5. Heat map of potentially interacting partner atoms for functional groups with sp^2 oxygen. For each detected partner atom, the distribution of in-plane (*x*-axis) and out-of-plane (*y*-axis) angles are shown. White dashed line indicates ideal geometry according to VSEPR model. Color scheme shows percentage of partner atoms.

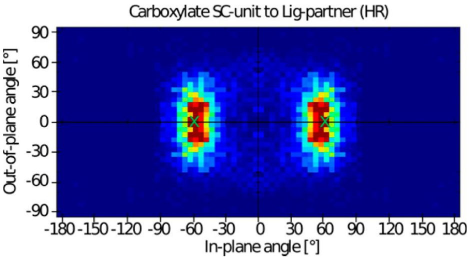


Figure 6. In-plane and out-of-plane angle distribution of carboxylate side chains. Image was generated by applying symmetry operations: green x = ideal interaction direction according to VSEPR model.

found in-plane of the lone pairs. Another study of the energetic dependency of the H-bond geometry, based upon a model system using formamide/formaldehyde, showed larger energy differences for out-of-plane deviations than for in-plane deviations.[34] Our analysis, however, suggests that exactly the contrary is the case, namely, that there are only small energetic differences for out-of-plane angles that broaden the distribution of out-of-plane angles, than for in-plane angles, that results in a narrower angle distribution.

Notable is also the primary amide distribution, O-Lig-unit to Any-partner unit, which shows a "side arm" of partner atoms with a high out-of-plane angle (40−60°) and an in-plane angle of 0−25° (see Figure 5e and Figure S6). Further analysis of the partners responsible for the "side arm" shows that the majority of these amides have an aromatic ring directly bound to the primary amide and most of the ligands in this group are nicotinamide adenine dinucleotide phosphate (NAD) derivatives (for an example, see Figure 7). The direct connection of the primary amide to the aromatic ring may cause a delocalization leading to a different tautomeric state of the primary amide group, which then consists of a hydroxyl group and a sp^2 iminyl nitrogen. In this case one would expect a tetrahedral arrangement of partner atoms around the oxygen instead of trigonal planar geometry. The example from Figure 7 additionally contains a water molecule that forms an H-bond with the NAD with an out-of-plane angle significantly deviating from the ideal value. Even though the distance to the amide

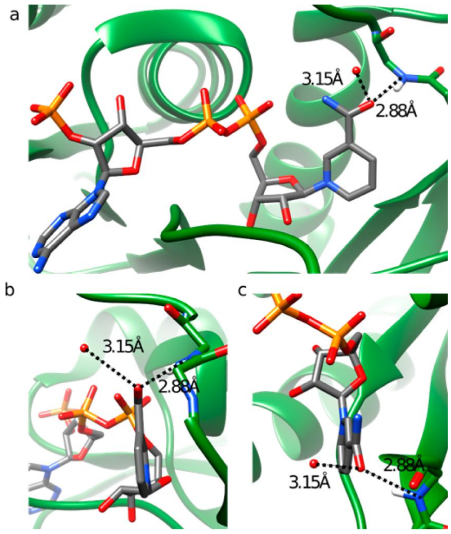


Figure 7. Example for special case of primary amide angle distribution, O-Lig-unit to Any-partner unit; atom pair between nicotinamide adenine dinucleotide phosphate and backbone nitrogen of serine 207-B: (a) overview, (b) side view of interaction pattern, (c) top view of interaction pattern with distinct out-of-plane deviation (PDB code 4cm3, resolution 1.95 Å; molecular graphics were created using UCSF Chimera[35]).

with 3.15 Å is rather large, suggesting a weak hydrogen bond, it further supports a potential tetrahedral geometry for the primary amide oxygen atom.

Angle Distribution at Hydroxyl Groups. Hydroxyl groups attached to a sp^3 hybridized carbon need to be differentiated from enols attached to an sp^2 carbon. According to the VSEPR model for sp^3 hybridized oxygen, a tetrahedral geometry would exhibit angles of about 70.5°, which correlates well with our statistical analysis (see Figure S7). According to the VSEPR model, enols should ideally have a trigonal planar geometry, which results in a 60° maximum in the angle histogram. However, our data show that enols that are either part of ligands or are side chains (tyrosine) have a slightly shifted maximum between 65° and 70° in the angular distribution (see Figure S7). This angle is more similar to a tetrahedral coordination. However, the point distributions around enol groups of side chains suggest two main interaction directions in plane with the aromatic ring (see Figure 8c,d), which resembles a sp^2 hybridization. The distribution around enols in ligand units on the other hand is less clear (see Figure 8a), but the main proportion of partner points is also located in plane with the aromatic ring (see Figure 8b, see higher (red) density areas). The more diverse distribution around enols in ligand units might be due to further substituents of the aromatic ring. Those have not been respected in this study but can lead to a difference in the polarity of the aromatic ring and lead to a tetrahedral geometry.

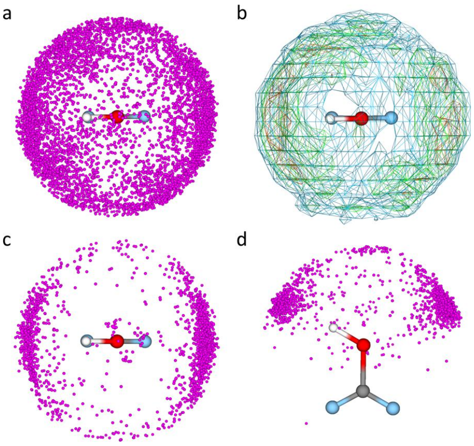


Figure 8. Partner atom (a) and density (b) distribution around enole groups Lig-unit to Any-partner unit: red = 50%, green = 25%, blue = 10% of maximum density value. (c, d) Partner atom distribution around enol groups SC-unit (=Tyr) to Lig-partner unit. Light blue atoms = linker.

Deeper analysis revealed that tyrosine side chains show higher deviations of up to 30° from the plane of the aromatic ring than enol groups in ligands (see Figure 8). These findings are similar to a previous study of intermolecular H-bonds of phenols.[36] In their study, the angle between the H-bond donor functionality and its H-bond partner deviates from the plane of the aromatic ring by up to 40° with even higher deviations for the acceptor depending on additional substitution of the aromatic ring.

Angle Distribution at Imidazole Nitrogen Atoms. The imidazole ring of histidine side chains can function as H-bond acceptor as well as donor depending on its tautomeric state. As expected from the VSEPR model, donor and acceptor density distributions have their main peak at 0° (see Figure 9). The donor distribution has a second peak around 50° in the case of the ε-nitrogen but not in case of the δ-nitrogen (see Figure 9b and Figure 10a,b). A further analysis revealed that these histidine side chains mostly coordinate metal ions (see Figure 10c,d). Additional atoms that coordinate the metal ion as well are also in close distance to the imidazole nitrogen, thus giving rise to the 50° peak. Therefore, this peak does not represent H-bond interactions and will be excluded from further conclusions drawn for H-bond geometries.

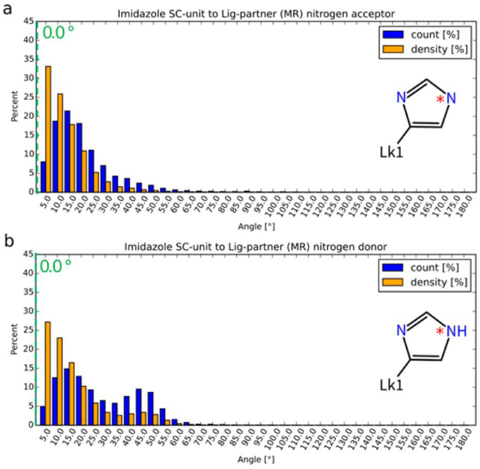


Figure 9. Imidazole (=His) partner angle histogram: (∗) histogram corresponds to the indicated nitrogen; green dashed line indicates ideal geometry according to VSEPR model.

Angle Distribution at Amide Nitrogen Atoms. The nitrogen atom of the amide functional group has a trigonal planar geometry. Due to the geometry definition according to the VESPR model, primary amides should have a distribution maximum in the frequency distribution at 60°, secondary amides at 0°. The nitrogen atom of tertiary amides cannot form any H-bonds.

The angle distributions in H-bonds between primary amides in protein side chains (N-SC-unit) and ligands (Lig-partner unit) and between secondary amides in ligands (N-Lig-unit) to any atoms (Any-partner unit) are in accordance with the VSEPR model (see Figure S8). However, the distribution derived for the interaction of primary amides in ligands with any atoms (N-Lig-unit to Any-partner unit) shows a deviation of 10° from the predicted ideal direction (see Figure 11a). A closer examination reveals that this shift is due to intramolecular interactions. They account for 17% of all partners of these primary amides and show a predominant preference for the anti-directed hydrogen of the nitrogen atom (see Figure S9).

Deviations from the predicted direction can also be observed for H-bonds between the secondary amides in the protein backbone (N-BB-unit) to ligands (Lig-partner unit, see Figure 11b). In the count and the density distribution, the peaks are

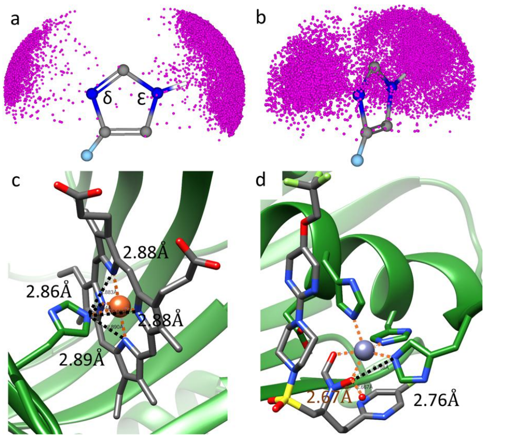


Figure 10. Special cases of imidazole partner distributions. (a, b) Molecule partner positions around imidazole in 2.6–2.9 Å distance. (b) Tilted side view on histidine side chain and the partner point distribution forming two main patches and a second ring around the ε-nitrogen in 35°–60° around the ideal interaction direction. (c) His 59-A complexes the iron of a heme group, and all nitrogen atoms are within 2.9 Å distance (PDB code 3tgc; resolution, 1.40 Å). (d) His 232-A complexes zinc 301-A and is within 2.76 Å distance of O1 of ligand AZ4, which already participates in two interactions, one to the metal and one to a water molecule (PDB code 4jp4; resolution, 1.43 Å); light blue atoms = linker (molecular graphics were created using UCSF Chimera[35]).

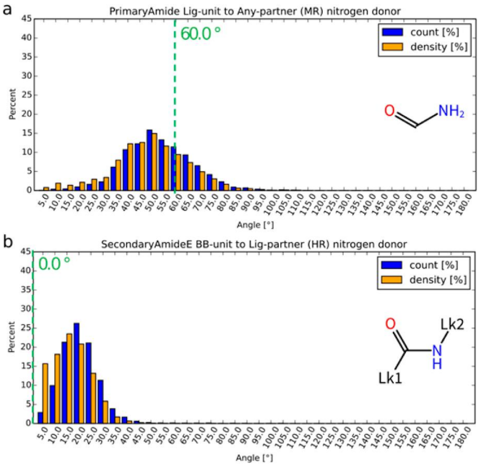


Figure 11. Amide partner angle histogram. Green dashed line indicates ideal geometry according to VSEPR model.

shifted by 15–20° from the expected peak at 0°. This observation is less prominent in the angle distribution of H-bonds between amides in ligands (N-Lig-unit) and any partner (Any-partner unit). Therefore, this observed shift might be due to spatial constraints from secondary structure elements as well as the orientation of close amino acid side chains (see Figure S10).

Angle Distribution at Guanidine Groups. The nitrogen atoms of the guanidine group can be further differentiated into sp^2 hybridized nitrogen atoms. Each sp^2 hybridized nitrogen would ideally have a trigonal planar geometry with bond angles of 120°. According to the VSEPR model, the maximum is expected at 60° and is indeed located between 60 and 65° (see Figure 12). A slight shift of density toward greater angles can be observed. The proximity of neighboring nitrogen atoms within the guanidine gives rise to this slight shift toward greater angles (see Figure 13). Similar observations were made in previous studies on H-bond geometries of protein side chains.[12]

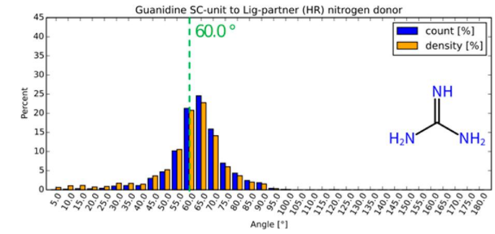


Figure 12. Guanidine partner angle histogram. Green dashed line indicates ideal geometry according to VSEPR model.

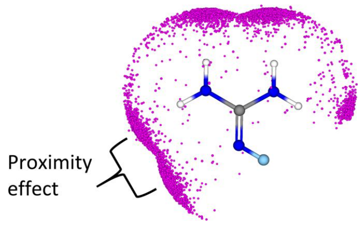


Figure 13. Examples for guanidine proximity effect; light blue atom = linker.

Angle Distribution at Ether Groups. The peak of the ether partner distribution would be expected to be around 55° assuming an ideal tetrahedral geometry. However, for the ether functional group the density distribution of partner points results in a continuous density with no separation of the two electron lone pairs (see Figure 14). Previous studies based on small molecule data found similar distributions.[11,15] However, no obvious reasons for this large difference to the expected lone pair directions could be found.

Angle Distribution at Amine and sp^2 Hybridized Nitrogen Groups. Depending on the hybridization state of the covalently bound atom, nitrogen atoms are expected to have a trigonal planar or tetrahedral geometry.

In the case of tetrahedral amines, partner atoms are expected around primary amines at 70.5° and around secondary amines at about 55°, both for donor and acceptor functions. In the case of tertiary amines, the partner atoms would ideally be around 0°. In case of a sp^2 hybridized, planar geometry, the nitrogen atom can only function as donor and the VSEPR model would predict partner acceptors at 60° for primary sp^2 hybridized nitrogen and at 0° for secondary sp^2 hybridized nitrogen. All histograms have their main peak at the expected angles (see Figure S11).

Additional information about the geometry of the interaction surface can be retrieved from the partner distribution of sp^2

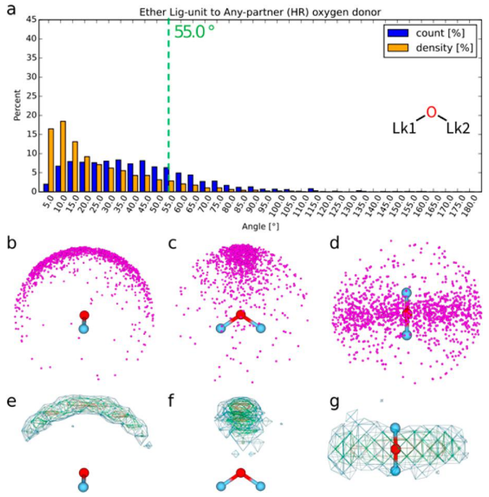


Figure 14. (a) Ether partner angle histogram; (b−d) partner points around ether; (e, f) partner density distribution (red = 50%, green = 25%, blue = 10% of maximum density value); light blue atoms = linker; green dashed line indicates ideal geometry according to VSEPR model.

hybridized nitrogen (see Figure 15). Tertiary amines and secondary sp^2 hybridized nitrogen only have a single acceptor or donor function, respectively. For both groups, about 90% of the partner atoms are within 35° from the ideal interaction direction (see Figure 4 and Figure 15a). A circular distribution of partner atoms can be observed around the interaction directions of primary sp^2 hybridized nitrogen (see Figure 4b). Additionally, the distributions are in accordance with the 35° deviations derived from tertiary amines and secondary sp^2 hybridized nitrogen. The deviation from the ideal acceptor direction (see Figure 4b) is highly similar to that of the imidazole acceptor (see Figure 10), with only slight differences in the absolute intensities.

Angle Distribution at Oxoacid Atom Pairs. In our data, all distributions of oxoacids have a maximum of partner contacts at angles between 55° and 65° as expected from the VESPR model (see Figure S12). Both sulfate and phosphate oxygens show a slight increase in the distribution between 0° and 15°. For phosphate oxygens, this effect is to a large extent due to coordination of a metal ion. Other atoms, which also coordinate the metal ion, are in close proximity to the phosphate oxygen (see Figure 16a). However, in the case of sulfate, a general explanation is more difficult to find. In many cases, the sulfate was integrated in extensive H-bond networks. In order to accommodate all surrounding donor functions, the deviations from the ideal interaction direction had to be less strict (see Figure 16b).

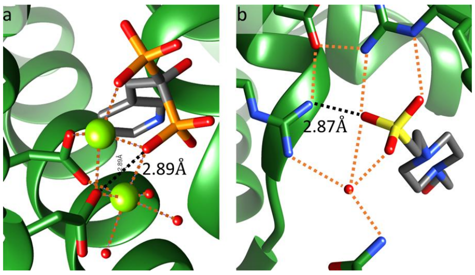


Figure 16. Example cases for phosphate and sulfate peaks between 0° and 15°. (a) Phosphate group coordinates metal (PDB code 4n9u; resolution, 2.11 Å). (b) Sulfate is integrated in extensive H-bond network (PDB code 1o28; resolution, 2.10 Å; molecular graphics were created using UCSF Chimera[35]).

Previous studies[37,38] based on CSD data observed smaller angles for the ideal interaction direction than the one observed here. The first study[37] defined the range of the interaction angles of oxoacids as between 30° and 70° (angle between P−O and either oxygen or nitrogen donor), and the second[38] observed a main direction between 30° and 35° (angle between P−O and hydrogen atom). While the deviation between our findings and the second study is clearly influenced by the considered atoms defining the angles, the difference to the first might show the influence of the data used for the analysis, i.e., within molecules versus protein−ligand interactions.

Sulfonamide had too few partner points around the oxygen atoms to draw meaningful conclusions. Therefore, only the partner points around the nitrogen atom were analyzed (see Figure S13). Their maximum is at the expected angle of 70.5°.

Influence of Resolution on the Hydrogen Bond Geometries. Two further evaluation steps were performed to investigate the influence of the quality of the X-ray data on the derived interaction geometries. First, the MR data set was divided into better than 1.5 Å and between 1.5 and 2.5 Å. Second, electron density was considered and only well-resolved functional groups and partner atoms were evaluated. Herein, the electron density of individual atoms (EDIA)[39,40] was used

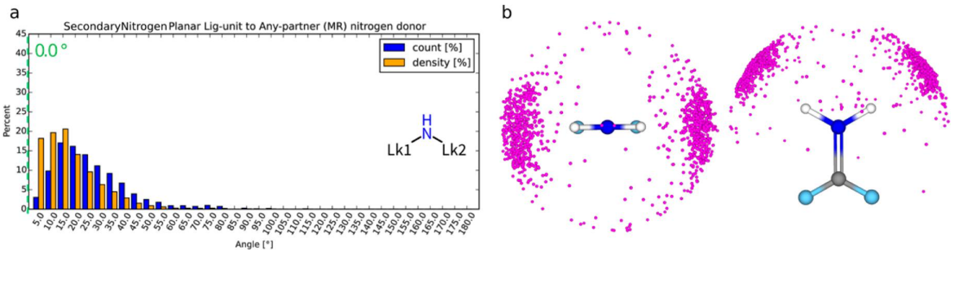


Figure 15. (a) Histogram of secondary sp^2 hybridized nitrogen. (b) Partner atom distribution of primary sp^2 hybridized nitrogen; green dashed line indicates ideal geometry according to VSEPR model; light blue atoms = linkers.

DOI: 10.1021/acs.jmedchem.7b00101

Table 5. Deviations for Different Interaction Geometries

		deviation		
functional group representatives	geometry	Optimum	First deviation	Maximum
Category: sp^2 Hybridized Oxygen Acc				
carboxyl	rectangle	in, 55°; out, 0°	in, 20°; out, 30°	in, 35°; out, 55°
ketone, amide (primary)	rectangle	in, 45°; out, 0°	in, 20°; out, 30°	in, 40°; out, 60°
amide (secondary, tertiary)	rectangle	in, 30°; out, 0°	in, 20°; out, 30°	in, 40°; out, 60°
ester	rectangle	in, 35°; out, 0°	in, 25°; out, 35°	in, 45°; out, 65°
Category: Generic Oxygen Acc				
ether,[a] ester (sp^3)[a,c]	rectangle	in, 0°; out, 0°	in, 10°; out, 40°	in, 25°; out, 70°
hydroxyl[a]	rectangle	in, 70.5°; out, 0°	in, 15°; out, [b]	in, 40°; out, [b]
hydroxyl (conjugated)[a]	rectangle	in, 60°; out, 0°	in, 15°; out, 30°	in, 40°; out, 50°
Category: Oxoacid Acc				
phosphate, sulfate, sulfonamide[c]	capped cone	55°	15°	40°
Category: Nitrogen Acc				
imidazole, nitrile,[c] amine (tertiary)	cone	0°	15°	40°
amine (secondary)	cone	55°	15°	40°
amine (primary)	cone	70.5°	15°	40°
Category: Generic Don				
amine, amide, guanidine, imidazole, hydroxyl,[c] hydroxyl (conjugated)	cone	0°	15°	40°

[a] The in-plane is defined by the C−O bond and the ideal interaction direction. [b] Out-of-plane angle deviation for freely rotatable hydroxyl oxygen acceptor cannot be derived from this study. [c] Functional groups with a statistically insignificant number of partner atoms were combined with most suitable ones.

to automatically differentiate between well resolved and less reliable functional groups and the partner atoms (EDIA cutoff: 0.8). This evaluation has been done for four different functional groups: (1) carboxylates, for which the electron lone pairs vary only slightly from the VSEPR expectations (Figure 5b and Figure 5c), (2) ketones, deviating by 15° from the expected electron lone pair directions (Figure 5a), (3) the nitrogen atoms of secondary amide (*E* isomer), with an average deviation for the donor direction of the nitrogen by 10° (Figure 11b), (4) ether oxygens, which have the highest average deviations from the expected electron lone pair directions (55°, Figure 14a).

The main interaction directions are not influenced by the resolution of the protein structures; i.e., specific geometric models used as input in the crystal structure refinement do not alter the distribution of partner atoms. This means that the deduced interaction geometries are not biased by the input of specific geometric models, though we expected that for lower resolution structures, the influence of the geometry models used in the refinement process might become visible. Overall, however, only minor differences in the angle distributions can be observed using either different resolution criteria or the EDIA criteria (see Figures S14−S17). Therefore, the derived interaction geometries are reliable and meaningful and do not show the influence of geometric models such as force field parameters used during structure refinement.

Statistically Derived Model for Interaction Geometries. The detailed analysis of partner distributions around functional groups allows the deduction of general geometric H-bond models.

For all interaction geometries a main interaction direction ("Optimum") was defined. Additionally, two threshold values were defined: first, the ideal angle deviation ("First deviation"), which includes about 65% of all partner points; second, the maximum angle deviation ("Maximum"), which covers about 95% of all partner points. H-bonds with an angle between "Optimum" and "First deviation" will be classified as ideal H-bonds. Those with an angle between "First deviation" and "Maximum" are distorted. If the angle deviates more than the "Maximum", we assume that no H-bond is formed. In addition to the angle deviations we also describe the geometric surface of the interaction direction, i.e., by a spherical cone, a capped cone, or a spherical rectangle (see Figure S18).

In order to reduce complexity, we tried to combine as many functional groups as possible for the definition of interaction geometries (see Table 5). Five main groups of interaction geometries were defined: (1) sp^2 hybridized oxygen acceptors, (2) generic oxygen acceptors, (3) oxoacid acceptors, (4) nitrogen acceptors, and (5) generic donors.

The sp^2-hybridized oxygen atoms were further divided into four subgroups since their partner atom distributions show distinct differences: (1) carboxyl, (2) ketones and primary amides, (3) secondary and tertiary amides, and (4) ester. Due to their different ranges of in- and out-of-plane angles, their interaction surface is best represented by a so-called spherical rectangle.

The group of "generic oxygen acceptors" includes the acceptor function of hydroxyl groups. Due to larger deviations of the enol group compared to the nonconjugated hydroxyls, the acceptor geometries are described separately. The sp^3 hybridized oxygen atom of the ester functional group had too few partner atom points to draw significant conclusions. That is in agreement with a theoretical and statistical analysis, which led to a similar observation, namely, that sp^3 oxygen atoms directly connected to an sp^2 carbon rarely form H-bonds.[41]

The distribution of interacting atoms around oxoacids is best described by a capped cone, i.e., a "halo" around the oxygen atom, as previously described by Bruno et al.[38]

A few other functional groups could not be statistically analyzed because too few partner points were identified. They are combined with those functional groups that best resemble their chemical features; i.e., the nitrile acceptor is described like any other nitrogen acceptor.

DOI: 10.1021/acs.jmedchem.7b00101

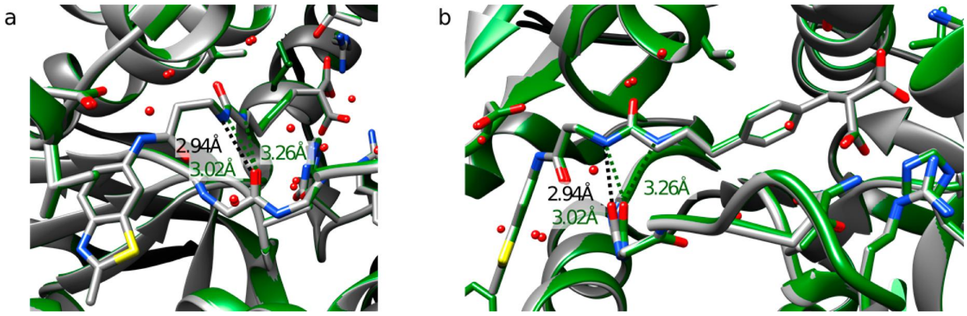


Figure 17. Example of a superimposed ligand pair that only deviates in one atom (change from amide, gray, PDB code 4ajo, to urea group, green, PDB code 4ajn). The interaction deviations for the nitrogen atoms of the urea group increase, while the binding affinity remains constant (amide 69 nM to urea 93 nM).[42]

CONCLUSION

The increase of available high-quality protein structures has allowed a new statistical evaluation of interaction geometries. In this study, 22 diverse functional groups and the preferred positions of potentially interacting partner atoms have been analyzed in detail. The importance of accurate H-bond geometry models can be observed in lactate dehydrogenase A (LDHA) compounds.[42] Here, the exchange of a carbon to a nitrogen leads to a second H-bond with nonideal geometry. Overall, this change does not lead to a significant change in binding affinity (Figure 17).

Lately "unconventional" H-bonds, like halogen bonds and weak H-bonds, e.g., C−H···O interactions, have drawn attention (for more information see reviews in refs 43−45). Due to the inferior strength of these weak interactions, they play only a minor role in the explanation of binding events.[46] Furthermore, due to the greater distances between the heavy atoms, they are less easily separable from van der Waals contacts.[28] Therefore, in our analysis we focused on the characteristics of strong H-bonds between free electron pairs of nitrogen or oxygen and hydrogen atoms covalently bound to nitrogen or oxygen. Comparison to the expectation of the widely accepted VESPR model and findings of previous studies showed both differences and similarities. In general, donor directions varied less often from the expected ideal interaction directions. However, electron lone pair directions deviate to a greater extent from the theoretical predicted directions. In particular, we observed significant deviation from the expected angle for sp^2 hybridized oxygen atoms. The deviation in out-of-plane direction was higher than in in-plane direction. Amines are in accordance with the VSEPR model, whereas oxoacids agreed more closely with a tetrahedral geometry than previously observed.[38]

H-bond angles of functional groups in proteins show more narrow distributions than the same functional groups in ligands. This suggests that functional groups in proteins are more restricted than those in ligands. The angle in the secondary amide from the protein backbone diverges more from the ideal angle than that from the Lig-unit. In order to fulfill its H-bond potential, it may need to adapt and build H-bonds with constrained interaction geometries. Furthermore, the distributions of partner atoms around functional groups not only represent the energetic prevalence but also contain adaptions due to steric hindrance. For example, a shift of the ideal interaction direction around nitrogen atoms of secondary amide BB-unit to Lig-partner unit might be due to close side chains. Another example is the hybridization of the oxygen atom of enols that might be affected by additional substituents of the aromatic ring.

Our analysis showed the geometric range of H-bond angles in protein−ligand structures, which would not be detectable using small molecule data only. Even though confined to the optimal H-bond length, unexpected geometries have been found, e.g., artificial peak for imidazole side chains due to metal interactions, great differences between sp^2 hybridized oxygen atoms. Most cases showed that such geometries are not caused by H-bonds but are mostly due to proximity effects, extended H-bond networks, or coordination to metals, which automatically gives rise to close neighbor contacts. If these geometries were to be included in computational models, e.g., for scoring protein−ligand interactions, then wrong or at least biased results could arise.

Even though the geometries were concluded from protein−ligand interactions, they are also valid for protein−protein and even ligand−ligand interactions (Figures S19−S22). Especially if backbone atoms are in close distance, they influence the distribution of partner atoms. Protein−protein interactions were, as expected at the beginning of this study, influenced by secondary structure element but also by short intra amino acid distances; i.e., the carboxylate group of the side chain is in close distance to the nitrogen donor of its backbone. However, the majority of the short distance protein−protein contacts is within the derived geometries and can thus be described and identified by them.

Especially for less well resolved structures, the refinement process has a greater influence on the final protein model. The influence of the refinement process was analyzed by splitting our data set into structures with high and medium resolution. Additionally, the electron density was taken into account as a quality criterion by calculating EDIA values for the functional group as well as its partner atoms. The angle distributions derived for a subset of functional groups (carbonyl, carboxyl, secondary amide, and ether) do not show significant differences using structures with different resolutions. Therefore, the H-bond geometries concluded from this study are indeed due to structural characteristics of the protein−ligand complexes.

From our comprehensive analysis of the geometric distribution of atoms around functional groups, general models for H-bond geometries can be concluded. These may assist further understanding of protein−ligand interactions and help guide the future development of computational models for protein−ligand interactions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jmedchem.7b00101.

Additional figures and tables (PDF)
Mol2 results of the used functional groups (TXT)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

ORCID

Eva Nittinger: 0000-0001-7231-7996
Agnes Meyder: 0000-0001-8519-5780
Matthias Rarey: 0000-0002-9553-6531

Present Address

[§]K.T.S.: OSTHUS GmbH, Eisenbahnweg 9-11, 52068 Aachen, Germany.

Author Contributions

[∥]E.N. and T.I. contributed equally.

Notes

The authors declare no competing financial interest.

An extended version of the tool that can be used for more detailed substructure analysis (NAOMInova) will soon be available at http://www.zbh.uni-hamburg.de/naominova.

■ ACKNOWLEDGMENTS

We thank our cooperation partner BioSolveIT, especially Markus Lilienthal, Holger Claussen, and Christian Lemmen, for helpful input. Special thanks are given to Hans Briem for fruitful discussions during the project. The project was part of the Biokatalyse2021 cluster and was funded by the Federal Ministry of Education and Research (BMBF) under Grant 031A183B.

■ ABBREVIATIONS USED

BB, backbone; CSD, Cambridge Structural Database; EDIA, electron density of individual atoms; FG, functional group; H-bond, hydrogen bond; HR, high- resolution; LDHA, lactate dehydrogenase A; Lig, ligand; MR, medium resolution; NAD, nicotinamide adenine dinucleotide phosphate; PDB, Brookhaven Protein Data Bank; rmsd, root-mean-square deviation; SC, side chain; VSEPR, valence shell electron pair repulsion

■ REFERENCES

(1) Pauling, L.; Corey, R. B.; Branson, H. R. The Structure of Proteins; Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proc. Natl. Acad. Sci. U. S. A.* **1951**, *37* (4), 205−211.

(2) Pauling, L.; Corey, R. B. The Pleated Sheet, a New Layer Configuration of Polypeptide Chains. *Proc. Natl. Acad. Sci. U. S. A.* **1951**, *37* (5), 251−256.

(3) Ramakrishnan, C.; Prasad, N. Study of Hydrogen Bonds in Amino Acids and Peptides. *Int. J. Protein Res.* **1971**, *3* (1−4), 209−231.

(4) Kroon, J.; Kanters, J. A.; van Duijneveldt-van De Rijdt, J. G. C. M.; van Duijneveldt, F. B.; Vliegenthart, J. A. O-H · O Hydrogen Bonds in Molecular Crystals a Statistical and Quantum-Chemical Analysis. *J. Mol. Struct.* **1975**, *24* (1), 109−129.

(5) Mitra, J.; Ramakrishnan, C. Analysis of O-H···O Hydrogen Bonds. *Int. J. Pept. Protein Res.* **1977**, *9* (1), 27−48.

(6) Vinogradov, S. N. Hydrogen Bonds in Crystal Structures of Amino Acids, Peptides and Related Molecules. *Int. J. Pept. Protein Res.* **1979**, *14* (4), 281−289.

(7) Ceccarelli, C.; Jeffrey, G. A.; Taylor, R. A Survey of O-H···O Hydrogen Bond Geometries Determined by Neutron Diffraction. *J. Mol. Struct.* **1981**, *70*, 255−271.

(8) Jeffrey, G. A.; Maluszynska, H. A Survey of Hydrogen Bond Geometries in the Crystal Structures of Amino Acids. *Int. J. Biol. Macromol.* **1982**, *4* (3), 173−185.

(9) Taylor, R.; Kennard, O.; Versichel, W. The Geometry of the N−H···O=C Hydrogen Bond. 3. Hydrogen-Bond Distances and Angles. *Acta Crystallogr., Sect. B: Struct. Sci.* **1984**, *40* (3), 280−288.

(10) Baker, E. N.; Hubbard, R. E. Hydrogen Bonding in Globular Proteins. *Prog. Biophys. Mol. Biol.* **1984**, *44* (2), 97−179.

(11) Murray-Rust, P.; Glusker, J. P. Directional Hydrogen Bonding to sp2- and sp3-Hybridized Oxygen Atoms and Its Relevance to Ligand-Macromolecule Interactions. *J. Am. Chem. Soc.* **1984**, *106* (4), 1018−1025.

(12) Ippolito, J. A.; Alexander, R. S.; Christianson, D. W. Hydrogen Bond Stereochemistry in Protein Structure and Function. *J. Mol. Biol.* **1990**, *215* (3), 457−471.

(13) Preissner, R.; Egner, U.; Saenger, W. Occurrence of Bifurcated Three-Center Hydrogen Bonds in Proteins. *FEBS Lett.* **1991**, *288* (1−2), 192−196.

(14) Sticke, D. F.; Presta, L. G.; Dill, K. A.; Rose, G. D. Hydrogen Bonding in Globular Proteins. *J. Mol. Biol.* **1992**, *226* (4), 1143−1159.

(15) Mills, J. E. J.; Dean, P. M. Three-Dimensional Hydrogen-Bond Geometry and Probability Information from a Crystal Survey. *J. Comput.-Aided Mol. Des.* **1996**, *10* (6), 607−622.

(16) Kortemme, T.; Morozov, A. V.; Baker, D. An Orientation-Dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein−Protein Complexes. *J. Mol. Biol.* **2003**, *326* (4), 1239−1259.

(17) Molcanov, K.; Kojić-Prodić, B.; Raos, N. Analysis of the Less Common Hydrogen Bonds Involving Ester Oxygen sp^3 Atoms as Acceptors in the Crystal Structures of Small Organic Molecules. *Acta Crystallogr., Sect. B: Struct. Sci.* **2004**, *60* (4), 424−432.

(18) Sarkhel, S.; Desiraju, G. R. N-H···O, O-H···O, and C-H···O Hydrogen Bonds in Protein-Ligand Complexes: Strong and Weak Interactions in Molecular Recognition. *Proteins: Struct., Funct., Genet.* **2004**, *54* (2), 247−259.

(19) Podtelezhnikov, A. A.; Ghahramani, Z.; Wild, D. L. Learning about Protein Hydrogen Bonding by Minimizing Contrastive Divergence. *Proteins: Struct., Funct., Genet.* **2007**, *66* (3), 588−599.

(20) Panigrahi, S. K.; Desiraju, G. R. Strong and Weak Hydrogen Bonds in the Protein-Ligand Interface. *Proteins: Struct., Funct., Genet.* **2007**, *67* (1), 128−141.

(21) Liu, Z.; Wang, G.; Li, Z.; Wang, R. Geometrical Preferences of the Hydrogen Bonds on Protein-Ligand Binding Interface Derived from Statistical Surveys and Quantum Mechanics Calculations. *J. Chem. Theory Comput.* **2008**, *4* (11), 1959−1973.

(22) Bilton, C.; Allen, F. H.; Shields, G. P.; Howard, J. A. Intramolecular Hydrogen Bonds: Common Motifs, Probabilities of Formation and Implications for Supramolecular Organization. *Acta Crystallogr., Sect. B: Struct. Sci.* **2000**, *56* (5), 849−856.

(23) Kuhn, B.; Mohr, P.; Stahl, M. Intramolecular Hydrogen Bonding in Medicinal Chemistry. *J. Med. Chem.* **2010**, *53* (6), 2601−2611.

(24) Gillespie, R. J.; Nyholm, R. S. Inorganic Stereochemistry. *Q. Rev., Chem. Soc.* **1957**, *11* (4), 339.

(25) Gillespie, R. J. Fifty Years of the VSEPR Model. *Coord. Chem. Rev.* **2008**, *252* (12−14), 1315−1327.

(26) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58* (3, Part 1), 380−388.

(27) 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 Res.* **2000**, *28* (1), 235−242.

(28) Desiraju, G. R. A Bond by Any Other Name. *Angew. Chem., Int. Ed.* **2011**, *50* (1), 52−59.

(29) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51* (12), 3199−3207.

(30) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32* (5), 922−923.

(31) Urbaczek, S.; Kolodzik, A.; Rarey, M. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *J. Chem. Inf. Model.* **2014**, *54* (3), 756−766.

(32) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. Close Agreement between the Orientation Dependence of Hydrogen Bonds Observed in Protein Structures and Quantum Mechanical Calculations. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (18), 6946−6951.

(33) Gillespie, R. J.; Robinson, E. A. Electron Domains and the VSEPR Model of Molecular Geometry. *Angew. Chem., Int. Ed. Engl.* **1996**, *35* (5), 495−514.

(34) Mitchell, J. B. O.; Price, S. L. The Nature of the N-H···O=C Hydrogen Bond: An Intermolecular Perturbation Theory Study of the Formamide/formaldehyde Complex. *J. Comput. Chem.* **1990**, *11* (10), 1217−1233.

(35) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera–a Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605−1612.

(36) Prout, K.; Fail, J.; Jones, R. M.; Warner, R. E.; Emmett, J. C. A Study of the Crystal and Molecular Structures of Phenols with Only Intermolecular Hydrogen Bonding. *J. Chem. Soc., Perkin Trans. 2* **1988**, No. 3, 265.

(37) Lommerse, J. P. M.; Taylor, R. Characterising Non-Covalent Interactions with the Cambridge Structural Database. *J. Enzyme Inhib.* **1997**, *11* (4), 223−243.

(38) Bruno, I. J.; Cole, J. C.; Lommerse, J. P.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. IsoStar: A Library of Information about Nonbonded Interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11* (6), 525−537.

(39) Nittinger, E.; Schneider, N.; Lange, G.; Rarey, M. Evidence of Water Molecules–a Statistical Evaluation of Water Molecules Based on Electron Density. *J. Chem. Inf. Model.* **2015**, *55* (4), 771−783.

(40) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments. Unpublished results, 2017.

(41) Böhm, H. J.; Klebe, G.; Brode, S.; Hesse, U. Oxygen and Nitrogen in Competitive Situations: Which Is the Hydrogen-Bond Acceptor? *Chem. - Eur. J.* **1996**, *2* (12), 1509−1513.

(42) Ward, R. A.; Brassington, C.; Breeze, A. L.; Caputo, A.; Critchlow, S.; Davies, G.; Goodwin, L.; Hassall, G.; Greenwood, R.; Holdgate, G. A.; Mrosek, M.; Norman, R. A.; Pearson, S.; Tart, J.; Tucker, J. A.; Vogtherr, M.; Whittaker, D.; Wingfield, J.; Winter, J.; Hudson, K. Design and Synthesis of Novel Lactate Dehydrogenase a Inhibitors by Fragment-Based Lead Generation. *J. Med. Chem.* **2012**, *55* (7), 3285−3306.

(43) Toth, G.; Bowers, S. G.; Truong, A. P.; Probst, G. The Role and Significance of Unconventional Hydrogen Bonds in Small Molecule Recognition by Biological Receptors of Pharmaceutical Relevance. *Curr. Pharm. Des.* **2007**, *13* (34), 3476−3493.

(44) Horowitz, S.; Trievel, R. C. Carbon-Oxygen Hydrogen Bonding in Biological Structure and Function. *J. Biol. Chem.* **2012**, *287* (50), 41576−41582.

(45) Ho, P. S. Biomolecular Halogen Bonds. *Top. Curr. Chem.* **2014**, *358*, 241−276.

(46) Etter, M. C. Hydrogen Bonds as Design Elements in Organic Chemistry. *J. Phys. Chem.* **1991**, *95* (12), 4601−4610.

Proteins*Plus*: a web portal for structure analysis of macromolecules.

[D6] Fährrolfes, R.; Bietz, S.; Meyder, A.; **Nittinger, E.**; Otto, T.; Flachsenberg, F.; Volkamer, A.; Rarey, M. Proteins*Plus*: a web portal for structure analysis of macromolecules. Nucleic Acids Res. 2017, 45 (W1): W337-W343.

https://doi.org/10.1093/nar/gkx333

Published online 2 May 2017 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* ***W337–W343***
doi: 10.1093/nar/gkx333

Proteins*Plus*: a web portal for structure analysis of macromolecules

Rainer Fährrolfes[1,†], Stefan Bietz[1,†], Florian Flachsenberg[1], Agnes Meyder[1], Eva Nittinger[1], Thomas Otto[1], Andrea Volkamer[2] and Matthias Rarey[1,*]

[1]Universität Hamburg, ZBH—Center for Bioinformatics, Bundesstrasse 43, 20146 Hamburg, Germany and [2]Institute of Physiology, Charité—Universitätsmedizin Berlin, Virchowweg 6, 10117 Berlin, Germany

Received January 31, 2017; Revised April 01, 2017; Editorial Decision April 14, 2017; Accepted April 18, 2017

ABSTRACT

With currently more than 126 000 publicly available structures and an increasing growth rate, the Protein Data Bank constitutes a rich data source for structure-driven research in fields like drug discovery, crop science and biotechnology in general. Typical workflows in these areas involve manifold computational tools for the analysis and prediction of molecular functions. Here, we present the Proteins*Plus* web server that offers a unified easy-to-use interface to a broad range of tools for the early phase of structure-based molecular modeling. This includes solutions for commonly required pre-processing tasks like structure quality assessment (EDIA), hydrogen placement (Protoss) and the search for alternative conformations (SIENA). Beyond that, it also addresses frequent problems as the generation of 2D-interaction diagrams (PoseView), protein–protein interface classification (HyPPI) as well as automatic pocket detection and druggablity assessment (DoGSiteScorer). The unified Proteins*Plus* interface covering all featured approaches provides various facilities for intuitive input and result visualization, case-specific parameterization and download options for further processing. Moreover, its generalized workflow allows the user a quick familiarization with the different tools. Proteins*Plus* also stores the calculated results temporarily for future request and thus facilitates convenient result communication and re-access. The server is freely available at http://proteins.plus.

INTRODUCTION

Three-dimensional (3D) structures of macromolecules are often the starting point for achieving an in-depth understanding of protein function. Their use has a long tradition in early-phase drug design applying tools like homology modeling, molecular docking and molecular dynamics simulation. Before any of these methods can be applied, the structure must be pre-processed and usually further analyzed. The preparation of a macromolecular model often includes the addition of hydrogen atoms, the identification of potential binding sites and the assembly of alternative conformations. While there have been substantial efforts of the worldwide Protein Data Bank (PDB) (1) to include information on the quality of deposited structures (2–5), additional validation of the atomic position reliability can be required for highly specific and more demanding applications. Visualization approaches are generally required for the analysis and interpretation of structural data and can further assist communication tasks like the illustration of molecular interactions. Other examples for advanced structure-based applications are the assessment of binding site druggability or the analysis of protein–protein interactions (PPI).

A wide range of tools has been developed to address these issues. However, the usability of these tools is occasionally restricted by platform dependencies, installation obstacles or non-trivial user interfaces. Especially command line tools might be challenging for non-expert users. Therefore, it is desirable to circumvent these issues by providing web services offering platform-independent usage and easy-to-use interfaces. For two of our own approaches, we already provided a web server (6,7). Both had their own interface fitting the specific requirements of the underlying methods. Thus, adding new functionalities or tools requires parallel refactoring or the development of a new web service. This does not only lead to a lack of interoperability but might also constitute a barrier for the users who need to familiarize themselves with different interfaces. In order to address these issues, we developed Proteins*Plus* which currently integrates the two former and four new state-of-the-art approaches. It also offers a unified, easy-to-use interface via a single web server. The integrated services cover a broad

[*]To whom correspondence should be addressed. Tel: +49 404 2838 7351; Fax: +49 40 42838 7352; Email: rarey@zbh.uni-hamburg.de
[†]These authors contributed equally to the paper as first authors.

range of elementary tasks frequently occurring in structure-related life sciences.

THE PROTEINS*Plus* SERVER

The main objective during the development of Proteins*Plus* was to create a general workflow to access and preprocess structural data for all kinds of life science research. The resulting workflow starts with the selection of a PDB ID or the upload of a custom PDB file and optionally a ligand file in SD format as input. Proteins*Plus* gives an immediate visual impression of the overall protein structure and contained ligand molecules. Afterward, the user can choose an application service of interest (see below), set additional tool configurations and start the calculation. The results will automatically be displayed after the calculation is finished. To provide the best possible user experience, Proteins*Plus* uses a caching system to store calculation results. With this system users can access results at a later time and share them with colleagues.

In order to allow for processing various kinds of structure-based tasks, a unified interface is needed that facilitates the integration of different services and meets high usability standards. The single main interface (cf. Figure 1) is divided into three panels and has a menu bar at the top to display additional target related information and to control the panels. The first panel visualizes 3D structural information with the NGL web viewer (8). Below is a control panel that allows to switch between different graphical representations, change the background color, display a molecular surface, clip the scene in z-direction and take a screenshot of the visualized data. If the given PDB file contains ligand molecules, these are additionally depicted as standard structure diagrams in the second panel and are further annotated with their PDB identifier and a unique SMILES string (9) (which is hidden per default). A click on a specific structural diagram highlights the ligand in the NGL viewer panel and also selects the ligand for the tool configuration. The third panel displays all tool related information and offers the ability to set options and trigger the calculations. After a calculation is finished, the result page will also be displayed in this panel. Depending on the applied tool, the result page contains various opportunities to manipulate the structure representation in the NGL viewer panel. This includes the visualization of calculated structural elements, the coloring of the depicted elements and the possibility to automatically focus on certain substructures. Linking the individual results with a commonly used 3D visualization supports the general understanding of different structural properties and simplifies the result interpretation.

Currently, the Proteins*Plus* server comprises six services addressing the most important tasks at the beginning of structure analysis. The following sections introduce the main aspects of these approaches.

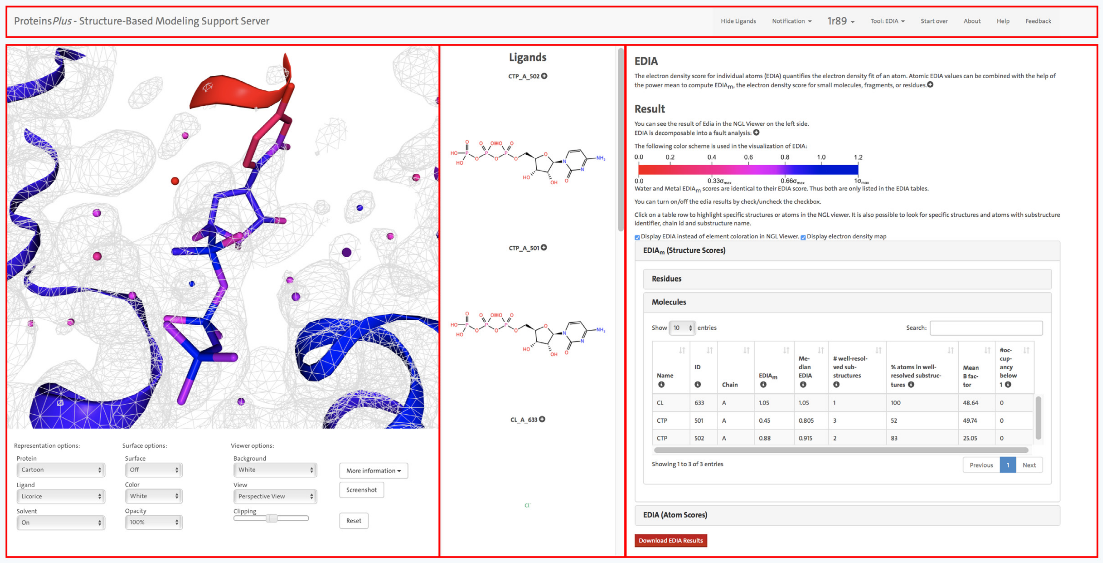
Protoss—hydrogen prediction

A common barrier to the application of three-dimensional structures is the incomplete representation of the respective macromolecules in many available data sources. This is primarily reasoned in shortcomings of the respective structure elucidation methods. For example, in the case of X-ray crystallography, insufficient resolution leads almost generally to the lack of hydrogen atom positions and frequently also impedes a differentiation of similar chemical elements which, in turn, increases the risk of erroneous side-chain orientations. Besides that, another common problem is the lack of additional information on bond orders and atom hybridization in many publicly distributed structural data sources. This is especially relevant for the interpretation of complexed ligands and atypical residues. However, a multitude of structure-based applications rely on a detailed representation of the considered molecules. For example, an accurate assessment of molecular interactions normally requires the knowledge of all atom positions, especially for the investigation of strongly directed interactions like hydrogen bonds. Therefore, several approaches have been developed for completing a structural model by missing elements such as hydrogen atoms and bond types and additionally improving unlikely side-chain orientations. (11–20)

The Proteins*Plus* server allows to tackle these tasks by applying our hydrogen prediction software Protoss (21,22). Starting with a macromolecular structure, Protoss first identifies unknown bond types on the basis of atom distance analysis. Following this, possible alternative states of polar moieties are detected and mutual energetic influences of these states are analyzed resulting in an interaction network. Finally, Protoss selects an optimal state for each group on the basis of a network optimization algorithm. The selected states eventually define the presence and position of polar hydrogen atoms as well as the orientation of ambiguous side chains. It is noteworthy that Protoss is able to consider alternative states of arbitrary chemical moieties (cf. Figure 2 for an example), while the vast majority of competitive tools focuses on the treatment of groups occurring in proteinogenic amino acids. Our large scale evaluation studies demonstrated that Protoss, in comparison to alternative approaches, benefits from this more elaborate modeling of chemical variability in terms of improved optimization capabilities for molecular interaction networks of protein–ligand interfaces. In the Proteins*Plus* web interface, the completed structures are visualized in the NGL viewer panel and provided for download in PDB format. Processed ligand molecules and atypical residues can additionally be downloaded in SD format. Due to its low computation times, the results of a Protoss calculation can mostly be provided within a few seconds.

PoseView—2D interaction diagrams

The increasing amount of protein–ligand complex structures—both from experimental sources and computational predictions—makes the availability of efficient visual inspection tools mandatory. The classic approach of inspecting such structure collections is looking at each of them in a 3D representation. This requires the user to rotate and translate the view until all features are visible. It can neither be used for the comparative visualization of many complexes nor for print and share. In text books and scientific publications, 2D representations which illustrate the key interactions between protein and ligand are frequently applied in this case. Various tools exist to condense the information about participating amino acids



Representation options:

Protein: Cartoon

Ligand: Licorice

Solvent: On

Surface options:

Surface: Off

Color: White

Opacity: 100%

Viewer options:

Background: White

View: Perspective View

Clipping

EDIA

The electron density score for individual atoms (EDIA) quantifies the electron density fit of an atom. Atomic EDIA values can be combined with the help of the power mean to compute EDIA$_m$, the electron density score for small molecules, fragments, or residues.

Result

You can see the result of Edia in the NGL Viewer on the left side.
EDIA is decomposable into a fault analysis:

The following color scheme is used in the visualization of EDIA:

0.0 0.2 0.4 0.6 0.8 1.0 1.2

0.0 0.33σ_{max} 0.66σ_{max} 1σ_{max}

Water and Metal EDIA$_m$ scores are identical to their EDIA score. Thus both are only listed in the EDIA tables.

You can turn on/off the edia results by check/uncheck the checkbox.

Click on a table row to highlight specific structures or atoms in the NGL viewer. It is also possible to look for specific structures and atoms with substructure identifier, chain id and substructure name.

☑ Display EDIA instead of element coloration in NGL Viewer. ☑ Display electron density map

EDIA$_m$ (Structure Scores)

Residues

Molecules

Show 10 entries Search:

Name	ID	Chain	EDIA$_m$	Median EDIA	# well-resolved substructures	% atoms in well-resolved substructures	Mean B factor	#occupancy below 1
CL	633	A	1.05	1.05	1	100	48.64	0
CTP	501	A	0.45	0.805	3	52	49.74	0
CTP	502	A	0.88	0.915	2	83	25.05	0

Showing 1 to 3 of 3 entries

EDIA (Atom Scores)

Figure 1. EDIA analysis for the crystal structure of an archaeal class I CCA-adding enzyme in complex with cytidine-5′-triphosphate (CTP) (PDB ID: 1R89 (10)). This figure demonstrates how the Proteins*Plus* web server can be used to assess the quality of a protein structure and analyze potential uncertainties in the structure. The panel on the right side shows the results from the EDIA calculation along with a short description of the quality measure. The detailed results for the EDIA$_m$ (structure score) for molecular substructures are displayed. For CTP 501 A, the EDIA$_m$ score is very low, indicating possible uncertainties in the structure. The binding site of this CTP molecule is shown in the left panel in the NGL web viewer, allowing a detailed visual inspection. All atoms in the structure are colored according to their individual EDIA score (as explained in the right panel). Additionally, the electron density map (*2fo-fc*) at 1 σ is displayed. It is clearly recognizable that most atoms in the cytosine moiety receive very low EDIA scores. This is consistent with the observation that around these atoms no electron density is observed at 1σ. The figure also highlights the menu bar at the top and all three panels with red rectangles, the NGL viewer with the control panel on the left, the ligand panel with structure diagrams in the middle and the tool panel with the result page of EDIA at the right.

and relevant interactions into a 2D structure diagram. MOE (24) and LeView (25) create diagrams that depict the ligand in atomic detail while residues of the pocket are shown as circles. LigPlot+ (26) and PoseView (27) show all interacting structural elements in atomic detail. Unlike LigPlot+, which generates 2D coordinates by flattening out the input 3D structure, PoseView generates structure diagrams from scratch focussing only on the best layout. Thus, it is able to draw about 80% of the Ligand Expo PDB subset without overlaps (28). Furthermore, PoseView aims at depicting all structure diagrams following the IUPAC drawing conventions. It is also integrated into the RCSB PDB website itself. An example of a PoseView diagram is given in Figure 2.

The Proteins*Plus* server facilitates to create PoseView interaction diagrams for ligands from PDB structures or additionally provided custom molecules in a fully automated fashion. Before identifying the involved amino acids, Protoss (see preceding section) is used for pre-processing the active site to define the protonation as well as tautomeric form of the protein and ligand. The resulting interaction diagram can be viewed directly in the browser and can be downloaded in various file formats (PDF, SVG and PNG).

EDIA—structural quality elucidation

Like any other experimental technique, structure elucidation has its limitations related to resolution and precision. Therefore, the examination of structural uncertainty is an advisable initial step for all applications based on macromolecular models. For structures determined with X-ray crystallography, a number of measures exist that objectively quantify the electron density fit, e.g. the real-space correlation coefficient (29) or the real-space difference density Z-score (30). Recently, we developed the electron density score for individual atoms (EDIA) (31) as a measure for estimating how well each atom position in a certain structure is supported by the experimental electron density. For all life scientists basing their research on individual structural features of a protein or a nucleic acid, it is essential to know this degree of experimental support for each atom, functional group or ligand molecule.

Based on a *2fo-fc* map, EDIA applies a grid-based approach to analyze the electron density distribution in a sphere around a certain atom considering both, density shape and intensity. It avoids the use of annotated B-factors by using a statistically determined resolution dependent B-factor. Therefore, EDIA overcomes known weaknesses of existing approaches like strong shape dependency (4) and tolerating overly flexible atoms that cause weak, stretched out electron density. The EDIA formula can be decomposed

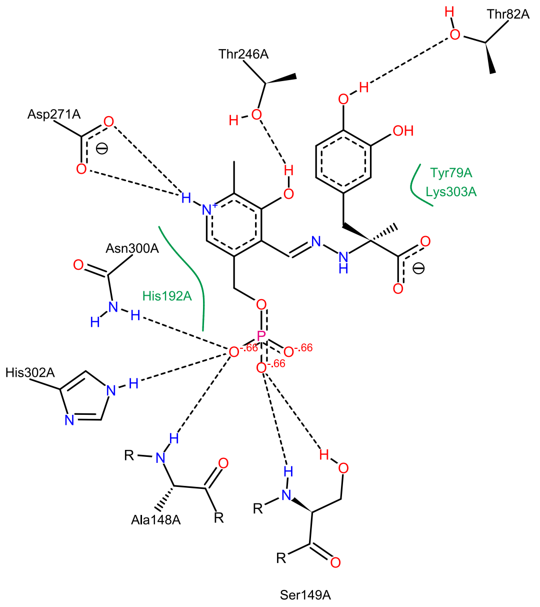


Figure 2. PoseView interaction diagram of dopa decarboxylase in complex with the inhibitor carbidopa (PDB ID: 1JS3 (23)). The automatically generated depiction clearly illustrates the molecular interactions described in the primary publication, e.g. the 'salt bridge between the carboxylate group of ASP 271 and the protonated pyridine nitrogen' (23). The PoseView interaction analysis is based on hydrogen orientations and protonation states calculated with Protoss (22).

to allow an automatic analysis explaining the reasons for a low EDIA score. Furthermore, EDIA scores can be combined using the power mean to score molecular fragments (EDIA$_m$) and thus facilitate the identification of well resolved substructures. EDIA$_m$ is also a very valuable addition to calculating RMSD values for investigation of re-docking capability, since the EDIA$_m$ truthfully reports the displacement from the experimental data while the RMSD reports the displacement from the interpreted coordinates.

Within Proteins*Plus*, EDIA and EDIA$_m$ scores are presented in an interactive table and the structure in the NGL viewer panel is recolored based on the EDIA coloring scheme. This allows an instantaneous differentiation of well resolved (blue) and weakly supported (red) substructures (see Figure 1). For comparison, the electron density can be displayed at a level of 1σ. Additionally, the result tables and the 3D visualization contain mutual links that allow to focus on a certain substructures in the NGL viewer panel by selecting an element from the result tables or filtering the entries of the result tables by clicking a certain residue in the viewer area. The download package consists of all EDIA and EDIA$_m$ scores in combination with the structure in a PDB file containing EDIA values in the B-factor column and the error analysis in the occupancy column. All EDIA scores of an average-sized structure can be computed in ∼4 min.

SIENA—structure ensemble assembly

When working with experimental structures of macromolecules, another highly relevant limitation is the inherent incapability of a single structure to properly represent the molecule's flexibility or other variations like its mutation sensitivity. As a straightforward approach to circumvent this drawback, multiple structures of the same target can be employed, often even without major adaption of the applied tools. Ideally, such ensembles can also be compiled from experimental data. While this remains difficult for nucleic acids, for which so far only a limited amount of refined structures exist, for many proteins there is already a sufficient number of structural alternatives available. The required ensemble generation process involves the challenge of selecting an appropriate set of structures. This includes the differentiation of desired and undesired variations as well as the identification of structural artefacts and inconsistencies in data annotation. Furthermore, typical preprocessing steps like a residue-wise alignment, superposition and hydrogen prediction (cf. Protoss) can support the direct applicability of the ensemble. In order to support all these tasks, we have developed an adaptive ensemble assembly approach called SIENA (32) that allows a case-specific generation and preprocessing of structure ensembles. Due to the high relevance of molecular interactions for protein functions, SIENA has a specific focus on the treatment of user-defined substructures like protein binding sites. SIENA achieves a quick access to alternative structures by a combination of an indexed database and an alignment technique (33) that is specifically geared to the processing of alternative binding site conformations. Additionally, it provides a set of various filters that allow a use-case specific adaption of the ensemble compilation. Among others, this includes functionalities for the assertion of structural consistency and an interaction-driven approach for ensemble reduction leading to a small but diverse set of representative structures. Various evaluation experiments highlight that SIENA allows for accurate and efficient ensemble preprocessing for sequence identities over than 70%.

Within the Proteins*Plus* server, SIENA can be triggered with a user-defined binding site query in combination with various filtering conditions to eliminate unwanted structures. Typical application scenarios like flexibility analysis, virtual screening and ligand pose comparison are supported by a one-click selection opportunity of predefined parameterization settings. The superimposed structures of the resulting binding site ensembles, which are usually provided within a few seconds, can be visualized in the NGL visualization area individually. Furthermore, the Proteins*Plus* server allows to download the generated ensemble in form of an archive file that contains all superimposed structures, a sequence alignment of the binding site residues and a statistical overview of certain ensemble measures like binding site RMSD or the number of mutated amino acids.

DoGSiteScorer—binding site detection

Target assessment is one of the major challenges in early drug discovery. Besides aspects such as medical rationale and commercial attractiveness, knowledge about the ability

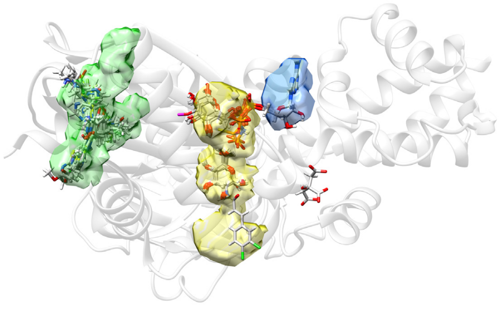


Figure 3. Predicted pockets using DoGSiteScorer for Hexokinase IV in complex with α-D-glucose only (PDB ID: 3QIC (34)). An ensemble was generated with SIENA using 3QIC as query structure and α-D-glucose as reference ligand. The figure includes all ligand molecules from this ensemble with more than six heavy atoms and within a distance of 5 Å from any protein atom in the 3QIC structure. As indicated by the superposition of ligands and DoGSiteScorer pocket predictions, the two best-ranked pockets correspond very well to the allosteric binding site (green) and the substrate binding site (yellow). Interestingly, the allosteric binding site is identified as the most druggable pocket (Drug-Score calculated by DoGSiteScorer: 0.81), which is in good agreement with the distribution of activating ligands found by SIENA. The ATP binding site, which is relatively solvent exposed, is not detected as one pocket but still well covered when considering the union of the two neighboring pockets depicted in yellow and blue.

of a target to bind a drug like molecule, i.e. called druggability, is of utmost importance (35). The binding site of a protein is the key to its function. Given a protein structure, the first step is, thus, the identification of potential cavities and a precise description of them. If a ligand-bound structure is available, this ligand defines the binding site. Nevertheless, additional allosteric or novel sites in ligand-free structures are of interest in prospective analyzes. In such cases, automatic methods to predict and rank cavities are investigated, e.g. FPocket (36), SiteMap (37) or DoGSiteScorer (38). Binding site detection methods rely solely on the 3D structure of the protein and use geometric and/or energetic information to detect cavities. Furthermore, these methods are able to estimate the druggable potential of a pocket using linear combinations (37), exponential functions (36) or machine learning models (38) derived from selected pocket descriptors, such as volume, enclosure or hydrophobicity.

DoGSiteScorer, is a grid-based pocket detection (39) and druggability prediction (38) method. The (sub)pocket detection step (39) has been evaluated on several benchmark dataset (Weisel dataset (40), PDBbind (41), sc-PDB (42)) and showed superior results. For druggability prediction (38), DoGSiteScorer uses a small set of physico-chemical and geometric descriptors combined with a support vector machine (SVM) trained and evaluated on the freely available druggability dataset (DD) (43). Validation on the complete DD yielded 88% correct predictions. DoGSiteScorer has been applied in several studies (>180 citations of references (7,38,39)) and was listed within the selected online resources supporting drug discovery in 2013 (44).

DoGSiteScorer is part of the Proteins*Plus* server and can be used to detect binding sites on a target of interest (see Figure 3). It discloses information about the properties of the detected pockets as well as their druggability. This knowledge can be used to prioritize targets for drug discovery or structures/binding sites for docking; or to compare pockets. As input, only a protein structure is required (PDB format or PDB ID). After pocket calculation, a sortable table appears that lists all pockets, together with the values for pocket surface, volume and druggability score. Additional descriptors can be displayed upon request. Per default, the largest pocket is shown in mesh representation in the NGL visualization (color corresponds to the table). Additional pockets can interactively be en-/disabled. All data, the pocket volumes (CCP4 format), the pocket residues (PDB format) as well as the full descriptor table (text format), is available for download.

HyPPI—protein–protein interactions classification

PPIs play key roles in biological regulatory pathways. Therefore, they are of central importance for the understanding of biological processes. Furthermore, they are of special interest for the development of small molecule modulators and lately received more attention in drug discovery (45–47). The PDB contains a substantial amount of structural data related to protein–protein complexes. However, the asymmetric unit (the smallest structure that cannot be recreated using symmetry operations) deposited in the PDB file is not necessarily composed of a biological-relevant protein–protein complex. The protein–protein complex might only be due to crystallization conditions (crystal artefact) or the biological-relevant complex must be generated by applying symmetry operations first. Since experimental methods for the determination of the oligomeric state of a complex are costly and time-consuming, it is of interest to develop an automated discrimination of biological complexes (permanent or transient) and crystal artefacts. Diverse methods exist which try to predict PPIs based on the computation of free energies or classification models based on physico-chemical and geometrical descriptors, e.g. PQS (48), NOXclass (49), EPIC (50), PISA (51), DiMoVo (52), CRK (53), OringPV (54), IPAC (55) or IChemPIC (56). Most of those methods achieve high accuracies of 85–97%. However, they use a large amount of descriptors to discriminate those complexes (22–213 descriptors).

The prediction tool HyPPI underlying Proteins*Plus* discriminates biological complexes and crystal artefacts. The most promising descriptors we found to characterize the different PPIs are the hydrophobic binding energy and the proportion of the interface ratios (IFquotient). The hydrophobic binding energy is calculated according to the desolvation term of the HYDE scoring function (57). The IFquotient measures the proportion of the subunits' relative interface area with respect to the molecular surface of the unbound subunit. Thus, it represents the symmetry of the PPI. Using only these two descriptors for the discrimination of biological complexes and crystal artefacts, we achieve a state-of-the-art accuracy of 92.5% on our training set of 254 complexes (49) and 77.9% on an independent test set (152 complexes from different sources (58–62)) which is comparable to the performance of the aforementioned tools. Within the Proteins*Plus* server, the discrimination of a PPI can be triggered with HyPPI by selecting the respective subunits. As a result, the probability for each class—biological (permanent or transient) versus crystal artefact—is given. This way, the user directly gets an indication of the reliability of the classification.

SUMMARY AND OUTLOOK

Proteins*Plus* presents a unified interface for various structure-based modeling tools. It makes the installation of large modeling software packages for an initial inspection of protein structural data dispensable. Therefore, the server is of special interest to life scientists with an occasional need to work with protein structures. The integrated NGL web viewer gives a first impression of the input structure and the calculated results. Thanks to the caching system, users can also share the results or check them later without any further calculation. With currently six tools, the unified easy-to-use interface and the generalized workflow, the Proteins*Plus* web server is a valuable resource for structure-based life science research. For the future, we plan to extend its functionality by additional modeling techniques and further improve its usability, e.g. by predefined use case parameterizations and by a pipeline functionality which allows to use previously calculated results as input for other integrated tools.

ACKNOWLEDGEMENTS

We would like to thank Alexander Rose for his support in integrating the NGL viewer on Proteins*Plus*. Figure 3 was generated with the UCSF Chimera package (63).

FUNDING

Developement of Proteins*Plus* by de.NBI (in part); German Federal Ministry of Education and Research [031L0105]. Funding for open access charge: German Federal Ministry of Education and Research (BMBF) [031L0105].

Conflict of interest statement. Protoss is part of the BioSolveIT product SeeSAR; M.R. is a co-founder and stakeholder of BioSolveIT.

REFERENCES

1. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980–980.
2. Adams,P.D., Aertgeerts,K., Bauer,C., Bell,J.A., Berman,H.M., Bhat,T.N., Blaney,J.M., Bolton,E., Bricogne,G., Brown,D. *et al.* (2016) Outcome of the first wwPDB/CCDC/D3R ligand validation workshop. *Structure*, **24**, 502–508.
3. Montelione,G.T., Nilges,M., Bax,A., Güntert,P., Herrmann,T., Richardson,J.S., Schwieters,C.D., Vranken,W.F., Vuister,G.W., Wishart,D.S. *et al.* (2013) Recommendations of the wwPDB {NMR} validation task force. *Structure*, **21**, 1563–1570.
4. Read,R.J., Adams,P.D., Arendall,W.B., Brunger,A.T., Emsley,P., Joosten,R.P., Kleywegt,G.J., Krissinel,E.B., Lütteke,T., Otwinowski,Z. *et al.* (2011) A new generation of crystallographic validation tools for the Protein Data Bank. *Structure*, **19**, 1395–1412.
5. Gore,S., Velankar,S. and Kleywegt,G.J. (2012) Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **68**, 478–483.
6. Stierand,K., Maaß,P.C. and Rarey,M. (2006) Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics*, **22**, 1710–1716.
7. Volkamer,A., Kuhn,D., Rippmann,F. and Rarey,M. (2012) DoGSiteScorer: a web-server for automatic binding site prediction, analysis, and druggability assessment. *Bioinformatics*, **28**, 2074–2075.
8. Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
9. Weininger,D. (1988) SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
10. Xiong,Y., Li,F., Wang,J., Weiner,A.M. and Steitz,T.A. (2003) Crystal structures of an archaeal class I CCA-adding enzyme and its nucleotide complexes. *Mol. Cell*, **12**, 1165–1172.
11. Brünger,A.T. and Karplus,M. (1988) Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins*, **4**, 148–156.
12. Bass,M.B., Hopkins,D.F., Jaquysh,W. A.N. and Ornstein,R.L. (1992) A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. *Proteins*, **12**, 266–277.
13. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
14. McDonald,I.K. and Thornton,J.M. (1995) The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. *Protein Eng.*, **8**, 217–224.
15. Hooft,R.W., Sander,C. and Vriend,G. (1996) Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins*, **26**, 363–376.
16. Word,J.M., Lovell,S.C., Richardson,J.S. and Richardson,D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**, 1735–1747.
17. Li,X., Jacobson,M.P., Zhu,K., Zhao,S. and Friesner,R.A. (2007) Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling. *Proteins*, **66**, 824–837.
18. Bayden,A.S., Fornabaio,M., Scarsdale,J.N. and Kellogg,G.E. (2009) Web application for studying the free energy of binding and protonation states of protein-ligand complexes based on HINT. *J. Comput. Aided Mol. Des.*, **23**, 621–632.
19. Labute,P. (2009) Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins*, **75**, 187–205.
20. Krieger,E., Dunbrack,R.L. Jr, Hooft,R.W. and Krieger,B. (2012) Assignment of protonation states in proteins and ligands: combining pKa prediction with hydrogen bonding network optimization. In: Baron,R (ed). *Computational Drug Discovery and Design*. Springer, NY, pp. 405–421.
21. Lippert,T. and Rarey,M. (2009) Fast automated placement of polar hydrogen atoms in protein-ligand complexes. *J. Cheminf.*, **1**, 13.
22. Bietz,S., Urbaczek,S., Schulz,B. and Rarey,M. (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J. Cheminf.*, **6**, 12.
23. Burkhard,P., Dominici,P., Borri-Voltattorni,C., Jansonius,J.N. and Malashkevich,V.N. (2001) Structural insight into Parkinson's disease treatment from drug-inhibited DOPA decarboxylase. *Nat. Struct. Biol.*, **8**, 963–967.
24. Clark,A.M. and Labute,P. (2007) 2D depiction of protein-ligand complexes. *J. Chem. Inf. Model.*, **47**, 1933–1944.
25. Caboche,S. (2013) LeView: automatic and interactive generation of 2D diagrams for biomacromolecule/ligand interactions. *J. Cheminform.*, **5**, 40.
26. Laskowski,R.A. and Swindells,M.B. (2011) LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.*, **51**, 2778–2786.
27. Stierand,K. and Rarey,M. (2007) From modeling to medicinal chemistry: automatic generation of two-dimensional complex diagrams. *Chemmedchem*, **2**, 853–860.
28. Stierand,K. and Rarey,M. (2010) Drawing the PDB: protein-ligand complexes in two dimensions. *ACS Med. Chem. Lett.*, **1**, 540–545.
29. Jones,T. and Kjeldgaard,M. (1997) [10] Electron density map interpretation. *Methods Enzymol.*, **277**, 173–208.
30. Tickle,I.J. (2012) Statistical quality indicators for electron-density maps. *Acta Crystallogr. D Biol. Crystallogr.* , **68**, 454–467.

31. Nittinger,E., Schneider,N., Lange,G. and Rarey,M. (2015) Evidence of water molecules—a statistical evaluation of water molecules based on electron density. *J. Chem. Inf. Model.*, **55**, 771–783.
32. Bietz,S. and Rarey,M. (2016) SIENA: efficient compilation of selective protein binding site ensembles. *J. Chem. Inf. Model.*, **56**, 248–259.
33. Bietz,S. and Rarey,M. (2015) ASCONA: rapid detection and alignment of protein binding site conformations. *J. Chem. Inf. Model.*, **55**, 1747–1756.
34. Liu,Q., Shen,Y., Liu,S., Weng,J. and Liu,J. (2011) Crystal structure of E339K mutated human glucokinase reveals changes in the ATP binding site. *FEBS Lett.*, **585**, 1175–1179.
35. Volkamer,A. and Rarey,M. (2014) Exploiting structural information for drug-target assessment. *Future Med. Chem.*, **6**, 319–331.
36. Le Guilloux,V., Schmidtke,P. and Tuffery,P. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.
37. Halgren,T.A. (2009) Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.*, **49**, 377–389.
38. Volkamer,A., Kuhn,D., Grombacher,T., Rippmann,F. and Rarey,M. (2012) Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.*, **52**, 360–372.
39. Volkamer,A., Griewel,A., Grombacher,T. and Rarey,M. (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.*, **50**, 2041–2052.
40. Weisel,M., Proschak,E. and Schneider,G. (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.*, **1**, 7.
41. Wang,R., Fang,X., Lu,Y. and Wang,S. (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, **47**, 2977–2980.
42. Kellenberger,E., Muller,P., Schalon,C., Bret,G., Foata,N. and Rognan,D. (2006) sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.*, **46**, 717–727.
43. Schmidtke,P. and Barril,X. (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.*, **53**, 5858–5867.
44. Villoutreix,B.O., Lagorce,D., Labbé,C.M., Sperandio,O. and Miteva,M.A. (2013) One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. *Drug Discov. Today*, **18**, 1081–1089.
45. Wells,J.A. and McClendon,C.L. (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, **450**, 1001–1009.
46. Ivanov,A.A., Khuri,F.R. and Fu,H. (2013) Targeting protein-protein interactions as an anticancer strategy. *Trends Pharmacol. Sci.*, **34**, 393–400.
47. Villoutreix,B.O., Kuenemann,M.A., Poyet,J.L., Bruzzoni-Giovanelli,H., Labbé,C., Lagorce,D., Sperandio,O. and Miteva,M.A. (2014) Drug-like protein-protein interaction modulators: challenges and opportunities for drug discovery and chemical biology. *Mol. Inform.*, **33**, 414–437.
48. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
49. Zhu,H., Domingues,F.S., Sommer,I. and Lengauer,T. (2006) NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, **7**, 27.
50. Block,P., Paern,J., Hüllermeier,E., Sanschagrin,P., Sotriffer,C.A. and Klebe,G. (2006) Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins*, **65**, 607–622.
51. Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.
52. Bernauer,J., Bahadur,R.P., Rodier,F., Janin,J. and Poupon,A. (2008) DiMoVo: a voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics*, **24**, 652–658.
53. Schärer,M.A., Grütter,M.G. and Capitani,G. (2010) CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. *Proteins*, **78**, 2707–2713.
54. Liu,Q. and Li,J. (2010) Propensity vectors of low-ASA residue pairs in the distinction of protein interactions. *Proteins*, **78**, 589–602.
55. Mitra,P. and Pal,D. (2011) Combining Bayes classification and point group symmetry under Boolean framework for enhanced protein quaternary structure inference. *Structure*, **19**, 304–312.
56. Da Silva,F., Desaphy,J., Bret,G. and Rognan,D. (2015) IChemPIC: a random forest classifier of biological and crystallographic protein-protein interfaces. *J. Chem. Inf. Model.*, **55**, 2005–2014.
57. Schneider,N., Lange,G., Hindle,S., Klein,R. and Rarey,M. (2013) A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: methods behind the HYDE scoring function. *J. Comput. Aided Mol. Des.*, **27**, 15–29.
58. Nooren,I. M.A. and Thornton,J.M. (2003) Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.*, **325**, 991–1018.
59. Bahadur,R.P., Chakrabarti,P., Rodier,F. and Janin,J. (2004) A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.*, **336**, 943–955.
60. De,S., Krishnadev,O., Srinivasan,N. and Rekha,N. (2005) Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct. Biol.*, **5**, 15.
61. Chen,Y.C. and Lim,C. (2008) Common physical basis of macromolecule-binding sites in proteins. *Nucleic Acids Res.*, **36**, 7078–7087.
62. Madaoui,H. and Guerois,R. (2008) Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 7708–7713.
63. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

From cheminformatics to structure-based design: Web services and desktop applications based on the NAOMI library.

[D7] Bietz, S.; Inhester, T.; Lauck, F.; Sommer, K.; von Behren, M. M.; Fährrolfes, R.; Flachsenberg, F.; Meyder, A.; **Nittinger, E.**; Otto, T.; Hilbig, M.; Schomburg, K. T.; Volkamer, A.; Rarey, M. From cheminformatics to structure-based design: Web services and desktop applications based on the NAOMI library. Journal of Biotechnology. J. Biotechnol. 2017, 261: 207-214.

https://doi.org/10.1016/j.jbiotec.2017.06.004

Journal of Biotechnology 261 (2017) 207–214

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Review

From cheminformatics to structure-based design: Web services and desktop applications based on the NAOMI library

Stefan Bietz[a], Therese Inhester[a], Florian Lauck[a], Kai Sommer[a], Mathias M. von Behren[a], Rainer Fährrolfes[a], Florian Flachsenberg[a], Agnes Meyder[a], Eva Nittinger[a], Thomas Otto[a], Matthias Hilbig[a], Karen T. Schomburg[a,1], Andrea Volkamer[b], Matthias Rarey[a,*]

[a] *Universität Hamburg, ZBH – Center for Bioinformatics, Bundesstrasse 43, 20146 Hamburg, Germany*
[b] *Institute of Physiology, Charité – Universitätsmedizin Berlin, Virchowweg 6, 10117 Berlin, Germany*

ARTICLE INFO

Keywords:
Cheminformatics
Structure-based design
Proteins*Plus*
NAOMI

ABSTRACT

Nowadays, computational approaches are an integral part of life science research. Problems related to interpretation of experimental results, data analysis, or visualization tasks highly benefit from the achievements of the digital era. Simulation methods facilitate predictions of physicochemical properties and can assist in understanding macromolecular phenomena. Here, we will give an overview of the methods developed in our group that aim at supporting researchers from all life science areas. Based on state-of-the-art approaches from structural bioinformatics and cheminformatics, we provide software covering a wide range of research questions. Our all-in-one web service platform Proteins*Plus* (http://proteins.plus) offers solutions for pocket and druggability prediction, hydrogen placement, structure quality assessment, ensemble generation, protein–protein interaction classification, and 2D-interaction visualization. Additionally, we provide a software package that contains tools targeting cheminformatics problems like file format conversion, molecule data set processing, SMARTS editing, fragment space enumeration, and ligand-based virtual screening. Furthermore, it also includes structural bioinformatics solutions for inverse screening, binding site alignment, and searching interaction patterns across structure libraries. The software package is available at http://software.zbh.uni-hamburg.de.

1. Introduction

Many biological and medicinal research questions highly benefit from the insights given by protein structure elucidation. Structural biology plays a key role in understanding, utilization, and manipulation of protein function. Therefore, the steadily increasing amount of publicly available structures in the Protein Data Bank (PDB) (Berman et al., 2000) constitutes a highly valuable resource for structure-based research. Structural bioinformatics contributes to this field with manifold powerful approaches. Computational methods are involved in structure elucidation, analysis, and quality assessment (Arzt et al., 2005; Goldsmith-Fischman and Honig, 2003; Kleywegt, 2000). Furthermore, they facilitate structure visualization, comparison, and the prediction of macromolecular properties. Preprocessing tools can be used to find appropriate data, to complete structures by missing elements, or to derive knowledge from the structural data that is needed for subsequent applications. Computational simulations like molecular dynamics, docking, and free-energy approximation support the understanding of physiological effects and aim to reduce the amount of necessary but expensive experimental analyses (Leach, 2001). In a similar manner, approaches from cheminformatics assist research on small molecules in areas like medicinal chemistry or biotechnology in general (Gasteiger and Engel, 2006). They support essential data management tasks like the identification of identical compounds, the storage in chemical databases, or filtering by molecular properties. Further applications are file conversion, pattern recognition in sets of similar molecules, or the enumeration of alternative conformations, tautomers, and protonation states. Cheminformatics also covers applications with a more predictive character. Examples are the generation of novel molecules (de novo design) (Schneider and Fechner, 2005) and the prediction of bioactive molecules (ligand-based virtual screening) (Koeppen et al., 2011). Several research questions in life sciences appear exactly at the interface of these two areas of computational science. Structure-based design is one of the key tools in early-phase drug and agrochemical discovery. Also, the development of novel techniques in biocatalysis benefits from this approach (Schneider et al., 2016). In

[*] Corresponding author.
E-mail address: rarey@zbh.uni-hamburg.de (M. Rarey).
[1] Present address: OSTHUS GmbH, Eisenbahnweg 9-11, 52068 Aachen, Germany.

http://dx.doi.org/10.1016/j.jbiotec.2017.06.004
Received 4 March 2017; Received in revised form 31 May 2017; Accepted 7 June 2017
Available online 11 June 2017

Table 1
Summary of all presented NAOMI-based web services and stand-alone applications.

Web services	Function	Main reference
DoGSiteScorer	Predicts the location of binding pockets and estimates their druggability.	Volkamer et al. (2012)
EDIA	Assesses the conformity of structural atom positions with the experimental electron density.	Nittinger et al. (2015)
HyPPI	Indicates whether protein–protein interactions are permanent, transient, or due to crystallization artefacts.	
PoseView	Draws 2D interaction diagrams of protein–ligand interactions.	Stierand et al. (2006)
Protoss	Adds hydrogen atoms to a macromolecular structure and optimizes their position with respect to polar interactions.	Bietz et al. (2014)
SIENA	Searches alternative binding site structures within the PDB.	Bietz and Rarey (2016)

Stand-alones	Function	Main reference
ASCONA	Calculates alignments of protein binding site conformations.	Bietz and Rarey (2015)
FSees	Enumerates novel molecules from a molecular fragment library.	Lauck and Rarey (2016)
MONA	Facilitates visualization and interactive processing of large molecular data sets.	Hilbig et al. (2013)
mRAISE	Identifies similar molecules via a ligand-based virtual screening approach.	von Behren et al. (2016)
PELIKAN	Searches user-defined protein–ligand interaction patterns in large structural databases.	Inhester et al. (2017)
SMARTSeditor	Supports an interactive design of SMARTS patterns.	Schomburg et al. (2013)
UNICON	Facilitates automatable coordinate generation, sampling of tautomers, protonation states and conformations as well as conversion of different file formats for small organic compounds.	Sommer et al. (2016)

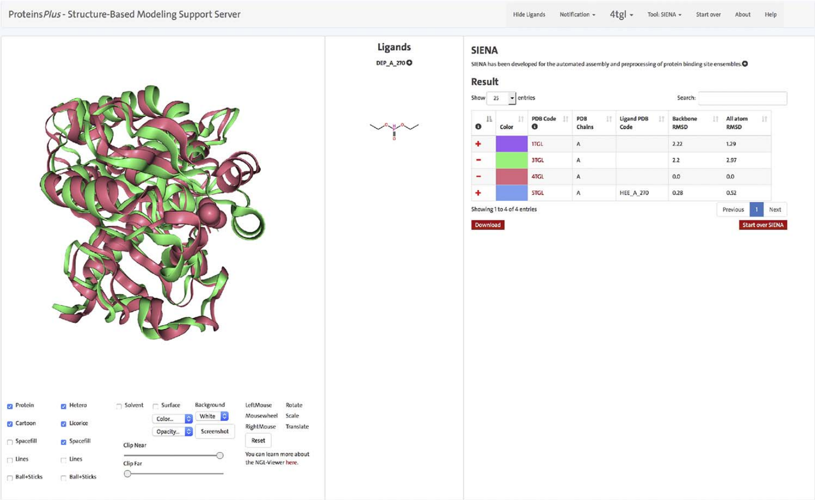
basic research, fields like chemical genomics and metabolomics show strong relationships to both fields. Based on a unique software infrastructure named NAOMI (Urbaczek et al., 2011, 2013, 2014), we have developed a wide range of methods that target problems related to chemistry and structural biology. The NAOMI software library supports cheminformatics and structural biology alike making it an ideal platform for the analysis of protein–ligand complexes. The software tools that we have built on the basis of these methods assist researchers from all areas of life science. Depending on their concrete application range and the respective user community, we either provide the tools as stand-alone software or in context of our web server Proteins*Plus*. In the following, we will briefly introduce all available applications grouped by their implementation strategy and scientific area. An overview of all web services and stand-alone applications is given in Table 1.

2. Web services

Providing the functionality of computational approaches via web services has various advantages over classical stand-alone approaches. Web services are usually the method of choice for providing access to large amounts of preprocessed data. They are platform independent, circumvent installation issues, and are thus accessible to the vast majority of the scientific community. Furthermore, web services mostly employ reduced, easy-to-use interfaces and therefore achieve higher usability and a more intuitive application behavior. For these reasons, we offer web-based solutions for many structure-related research questions. In order to support quick familiarization with the supplied functionality, all of our services are integrated into a single web platform, called Proteins*Plus* (Fährrolfes et al., 2017), offering a unified interface and a standardized workflow for all featured applications. Proteins*Plus* can either operate on a PDB structure (by providing the PDB ID) or on files uploaded by the user (PDB format for macromolecules, SD format for small molecules). The provided structure is visualized as a three-dimensional model using the NGL viewer (Rose and Hildebrand, 2015) (cf. Fig. 1). Its integration into Proteins*Plus* allows several control options including various depiction styles for both protein and ligands, surface visualization, and screenshot generation. The 3D window is also used to illustrate the results for most of the tools integrated into Proteins*Plus*, e.g. binding pockets, predicted hydrogen positions, or electron density fit. Based on a molecule perception algorithm (Urbaczek et al., 2013), ligand molecules are additionally depicted as structure diagrams and SMILES strings. Textual results are presented as sortable tables and can be downloaded for further processing.

Proteins*Plus* covers solutions for multiple problems like structure preprocessing, analysis, and visualization issues as well as the prediction of macromolecular properties. One of the most essential tasks in the context of structure-based research questions is the assessment of the structure's quality. Due to experimental uncertainties and modeling inaccuracies, there might be less experimental evidence for certain parts of a structural model. As this affects the reliability of subsequent interpretations and calculations, it is of great importance for many structure applications to identify such structural uncertainties. In the case of X-ray crystallography, some of these potential error sources can be detected by comparing the experimental electron density with the derived structural model. Various measures have been developed that quantify differences of experimental and modeled structure representations (Jones and Kjeldgaard, 1997; Tickle, 2012) albeit with slightly different purposes. Our recently developed electron density score for individual atoms (EDIA) (Nittinger et al., 2015) aims at the identification of structural elements insufficiently supported by experimental data. In contrast to other methods, this also includes highly flexible substructures, although different uncertainties are still captured with a single consistent measure. Furthermore, EDIA facilitates an atom-wise quality description which, e.g., can be used for an intuitive graphical representation (as included in Proteins*Plus*) or the exclusion of unreliable atoms from conformation-critical analysis strategies.

Another common problem is that most protein structures do not provide a full and precise model of all atoms. Usually, this is due to certain drawbacks of the approaches applied for structure elucidation. For X-ray crystallography, the major issues are the identification of hydrogen positions and the identification of certain side chain orientations, which are both often complicated by insufficient resolution (Davis et al., 2003, 2008). Additionally, numerous structures also lack detailed information on bond orders of atypical residues and ligand molecules. However, many applications dealing with the assessment of molecular interaction like binding affinity estimation or molecular dynamics simulations rely on a complete and accurate atomistic model of the protein. Our hydrogen prediction approach Protoss (Lippert and Rarey, 2009; Bietz et al., 2014) can be used to complete a given structural model by hydrogen atoms, assign unknown bond types, and correct erroneous side chain orientations. In order to achieve an optimal orientation of hydrogen atoms, Protoss optimizes the orientation of rotatable hydrogen atoms and considers alternative protonation states in both ligand and protein moieties. In contrast to this, most competing tools (Brünger and Karplus, 1988; Bass et al., 1992; McDonald and Thornton, 1994, 1995; Hooft et al., 1996; Word et al., 1999; Li et al., 2007; Bayden et al., 2009; Labute, 2009; Krieger et al., 2012) mainly handle those functional groups existing in proteins while neglecting the majority of groups occurring in ligand molecules. This



SIENA

SIENA has been developed for the automated assembly and preprocessing of protein binding site ensembles.

Result

Show 25 entries Search:

	Color	PDB Code	PDB Chains	Ligand PDB Code	Backbone RMSD	All atom RMSD
+	🟪	1TGL	A		2.22	1.29
−	🟩	3TGL	A		2.2	2.97
−	🟥	4TGL	A		0.0	0.0
+	🟦	5TGL	A	HEE_A_270	0.28	0.52

Showing 1 to 4 of 4 entries

You can learn more about the NGL-Viewer here.

Fig. 1. The Proteins*Plus* server. The right panel shows an ensemble of triacyl-glycerol acylhydrolase calculated with SIENA (Bietz and Rarey, 2016). Two alternative conformations from this ensemble are depicted in the left panel using the NGL viewer (Rose and Hildebrand, 2015). The middle panel illustrates the ligand from the query structure.

more precise modeling allows Protoss to reduce the number of undesirable atom contacts in the binding site and reaches a better agreement with manually assigned protonation states of ligand molecules.

Besides missing or misplaced atoms, another major drawback is that single structures determined by crystallographic approaches do not properly represent important aspects of a macromolecule's nature like its structural flexibility or its sensitivity to mutations. A straightforward manner to consider and investigate these aspects is to use a set of alternative structures of a certain target. Still, a reasonable application of such structural ensembles requires a careful ensemble generation. With SIENA (Bietz and Rarey, 2016), the Proteins*Plus* server contains a powerful tool for ensemble generation that can be applied for various use cases. SIENA detects all relevant structures from the PDB using ASCONA (Bietz and Rarey, 2015), an alignment algorithm specifically adapted to the treatment of alternative conformations. In contrast to predefined ensemble datasets (Verdonk et al., 2008; An et al., 2005; Kufareva et al., 2012), SIENA allows to adapt the ensemble assembly to meet the specific requirements of different ensemble applications. This is realized with the aid of various alignment configurations and additional property filters. For example, SIENA can be used to identify structural artefacts like missing atoms or chemically modified residues and, thus, to ensure the structural consistency of the ensemble, a property that is highly relevant for ensemble docking. It also includes algorithms that reduce the size of large ensembles to a small set of representative structures, covering either the overall conformational space or the more local variety of interaction capacities within the binding site.

In addition to the preprocessing tools described above, Proteins*Plus* also covers frequently required structure analysis tasks. Since the binding site of a protein is the key to its function, DoGSiteScorer (Volkamer et al., 2010, 2012) provides functionalities for automated protein pocket detection and druggability analysis. Both aspects are highly relevant for the assessment and prioritization of target proteins, especially in drug discovery projects. Pocket detection (Volkamer et al., 2012; Le Guilloux et al., 2009; Halgren, 2009) algorithms can be applied for the recovery of known binding sites as well as the identification of novel cavities, e.g. for allosteric modulators. They also lay the foundation for calculating pocket descriptors that allow property prediction or target comparison. DoGSiteScorer uses a purely structure-based approach for the prediction of potential binding sites. Pockets can be split into smaller subpockets, which allows a more granular pocket description and indicates potential ligand expandabilities. For all identified cavities it further derives a set of physicochemical and geometric descriptors, which can be consulted to gain deeper insights into a protein's functionality. Additionally, DoGSiteScorer applies these descriptors for druggability prediction using a support vector machine. The resulting druggability scores can be used to rank the identified pockets and select the most promising ones for further investigations. Within the Proteins*Plus* server, users can visualize the pockets predicted by DoGSiteScorer as an overlay with the 3D structure and sort them by their calculated features. An example is given in Fig. 2.

Another problem that requires predictive analysis is the assessment of protein–protein interactions (PPI). PPIs are the foundation of regulatory pathways and are therefore essential for the understanding of signal transduction processes (Pawson and Nash, 2003). Other prominent examples where protein–protein complex formations play a major role for protein function are transcription factors, oligomeric enzymes, and the composition of immunoglobulins. The involvement of these complexes in pathological phenomena also provokes the interest of

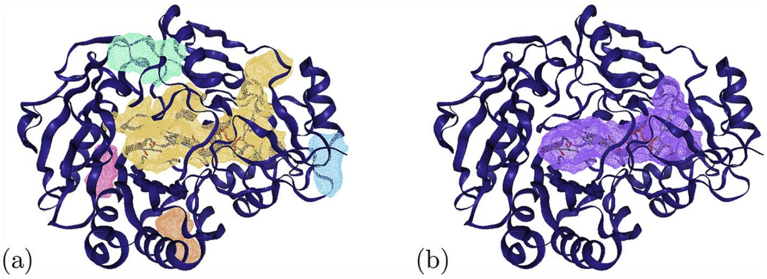


Fig. 2. Alditol oxidase in complex with sorbitol (PDB ID 2VFT Forneris et al., 2008). (a) The five largest pockets and (b) the best scored subpocket identified by DoGSiteScorer, here visualized as colored grids. (b) Shows that the largest subpocket (depicted in purple) is in good agreement with the binding mode of sorbitol and the cofactor FAD. The figure has been created with the NGL viewer (Rose and Hildebrand, 2015) as integrated in Proteins*Plus*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

drug design campaigns in targeting protein–protein interfaces (Wells and McClendon, 2007; Ivanov et al., 2013; Villoutreix et al., 2014). The PDB provides a solid basis for structure-based investigations of PPIs. In this context, it is necessary to differentiate between complexes that have a biological impact and those that occur only due to crystallization artefacts. Further, it is relevant to distinguish permanent and transient complexes, especially for inhibiting protein–protein interactions. Various approaches exist that can be applied for predicting such classifications (Henrick and Thornton, 1998; Zhu et al., 2006; Block et al., 2006; Krissinel and Henrick, 2007; Bernauer et al., 2008; Schärer et al., 2010; Liu and Li, 2010; Mitra and Pal, 2011; Da Silva et al., 2015). Within the Proteins*Plus* server the PPI prediction tool HyPPI is used for this task. It is based on an energy approximation using the desolvation term of the HYDE scoring function (Schneider et al., 2013) and the interface area ratio of both binding partners representing the interface symmetry. With only these two descriptors, HyPPI achieves a state-of-the-art accuracy for the discrimination of permanent, transient, and artificial complexes. The result is indicated by a probability value for each possible class that allows an additional assessment of the prediction precision.

Protein function is often strongly related to a protein's interaction with small molecules like substrates, activators, or inhibitors. Thus, a clear depiction of molecular interactions is essential for knowledge exchange. Since two-dimensional figures are still the conventional means for such tasks, there are two major alternatives: Following the first strategy, a three-dimensional model is projected onto a two-dimension layer. Although this is a more realistic representation, it also raises the necessity of choosing a tradeoff between depicting an adequate amount of information and preventing superfluous overlaps that impede an intuitive interpretation. The second approach illustrates the molecules using a two-dimensional simplification abstracting from the spatial constitution. A classic example of this concept is the early-introduced convention to depict molecules by 2D structure diagrams. In a similar manner, protein–ligand interfaces can be depicted by neglecting the overall spatial orientation of the participating moieties but still illustrating their main interactions. Several tools have been developed which address this very task (Clark and Labute, 2007; Laskowski and Swindells, 2011; Caboche, 2013). Our tool PoseView (Stierand et al., 2006; Stierand and Rarey, 2007) distinguishes itself from other approaches by depicting all moieties from the protein–ligand interface at atomic detail following IUPAC conventions. It optimizes the overall layout from scratch considering only the depicted molecular interactions between the ligand and the surrounding protein residues but uses no further spatial relations as geometric constraints. This way, it can draw about 80% of all interfaces for molecules from the LigandExpo dataset without graphical overlaps (Stierand and Rarey, 2010). Within Proteins*Plus*, PoseView can be applied to any structure from the PDB or, alternatively, to a custom protein–ligand complex provided by the user. PoseView figures are also used at the PDB webpage to describe the molecular interactions of a protein's ligands.

3. Future extensions to Proteins*Plus*

In the future, Proteins*Plus* will be developed further with the aim of providing a more comprehensive interface and augmenting its functionality. On the one hand, we plan to extend the Proteins*Plus* interface to allow partial integration into foreign websites. This will facilitate a more straightforward access to specific requests. On the other hand, we will integrate additional tools. In particular, we plan to support a web-based access to our protein binding site screening engines TrixP (von Behren et al., 2013) and iRAISE (Schomburg et al., 2014).

TrixP is a virtual screening technology for the comparison of protein binding sites. This task is relevant for problems like the classification of structurally resolved proteins with unknown function, identification of potential substrates, or the discovery of new inhibitors. The detection of similar binding sites can also indicate possible adverse effects of drugs or highlight opportunities for developing poly-pharmacological ligands. Because of its high biological relevance and broad application range, many approaches exist which address binding site comparison (see Kellenberger et al., 2008; Nisius et al., 2012; Jalencas and Mestres, 2013 for reviews). TrixP compares binding sites on the basis of abstracting descriptors (Schlosser and Rarey, 2009) encoding pharmacophore properties and their geometrical distances. Due to their discrete character, these descriptors can be efficiently searched using a bitmap index which, in turn, allows a much more efficient identification of similar binding sites compared to classical linear search routines. While alignment-free, fingerprint-based methods – which often also provide considerable speed-up – exhibit a loss of interpretability, TrixP still provides a structural alignment for the identified binding site matches and does therefore not suffer from this limitation. Furthermore, TrixP also accounts for a certain degree of protein flexibility by accepting partial shape similarity during the descriptor comparison. Compared to other state-of-the-art methods, the TrixP predictions exhibit mostly equivalent and partially superior results. The integration of TrixP into Proteins*Plus* will allow to search for similar binding sites across all known binding sites in PDB structures. This process can benefit from using the SIENA approach via identifying and eliminating equivalent binding sites for redundancy reduction.

Another strategy that tackles phenomena like adverse effects, poly-pharmacology, and target-specificity from a different perspective is protein target prediction. Starting with a certain ligand molecule, the problem to be solved is the detection of all proteins possibly complexing the ligand. In order to avoid expensive experimental analyses for obtaining a molecule's target profile, computational methods can be applied (Keiser et al., 2007; Hopkins, 2008; Campillos et al., 2008; Rognan, 2010). Our inverse screening protocol iRAISE (Schomburg et al., 2014) addresses this task with a structure-based procedure. It employs similar pharmacophore feature-based descriptors and index-based search technique as TrixP (see above). In order to deal with noise observed for scoring functions when applied to different targets (Wang et al., 2012), iRAISE employs a tailor-made scoring cascade. Several normalization approaches with respect to the score of a structure's co-crystallized ligand and the mutual coverage of target pocket and placed ligand are included. By this means iRAISE was able to outperform a classical docking protocol applied to target prediction. Furthermore, its structure-based strategy has the additional advantage of providing binding mode hypotheses. This might help to understand modes of action and highlight opportunities for lead optimization. A large-scale evaluation also showed that iRAISE exhibits superior performance compared to a sequence-based method for off-target prediction in the case of more distantly related targets, while the latter exhibits a better enrichment for ligands that bind to proteins with higher sequence identity. Thus structure- and sequence-based methods might be used in combination as they complement each other.

4. Desktop applications

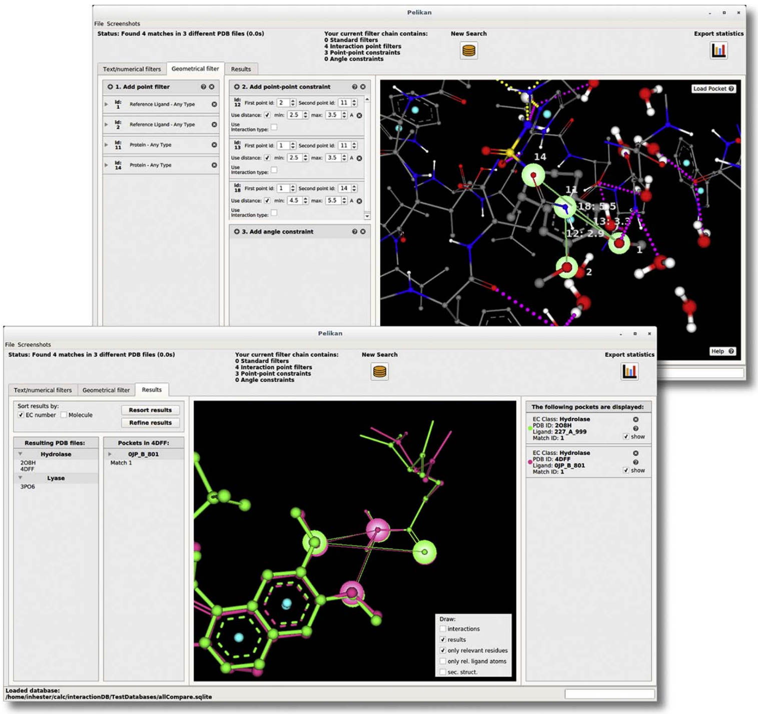
Compared to web services, desktop applications might be the favorable solution for certain methods that require a more comprehensive interface or shall be incorporated in fully automated workflows. Using in-house hardware can also be of advantage if an application requires high data traffic or the user's input data is confidential. For these reasons, we also provide many of our approaches as stand-alone software solutions. In order to reduce technical barriers, all of our tools can be directly used after downloading and unzipping the respective package. Additional installation is possible and might improve personal work habits but is not required for any application. Most applications are available for Windows, Linux, and macOS. The collection of software tools is comprised of several approaches targeting problems from structural biology. Users that prefer a local tool set-up can work with the stand-alone executables of iRAISE for inverse screening and target profiling and SIENA for ensemble generation (cf. Section 2).

Additionally, we also offer a stand-alone software package of the alignment approach ASCONA (Bietz and Rarey, 2015), which is applied by SIENA for the alignment of alternative binding site conformations. ASCONA has a strong focus on the alignment of alternative conformations and the detection of substructures like binding sites rather than aligning the whole protein, although this is also supported. Concerning these purposes, most other alignment techniques have the disadvantage that they either rely on detecting similar structures, which is obviously counter-productive when dealing with alternative conformations, or they apply only sequence-based descriptors but neglect structural features, which makes it impossible to distinguish multiple copies of the same peptide chain in homo-oligomeric proteins. However, this is often necessary for obtaining a distinct residue assignment if the binding site is formed by the interface of different subunits. ASCONA combines the benefits of both approaches and thus constitutes a unique solution for aligning binding site conformations. Besides that, it also supports other residue-wise alignment and structural superposition tasks as long as the input structures exhibit a sequence similarity of at least 70%.

The investigation of interaction patterns in protein–ligand interfaces is of great interest to drug discovery. Typical tasks for this application are the detection of specific embeddings of molecular substructures in their chemical environment. Moreover, interaction preferences of specific molecular substructures can be revealed. The tool PELIKAN (Inhester et al., 2017) can be used to rapidly mine large sets of protein–ligand interfaces for specific geometrical arrangements of atoms. The underlying method goes beyond other existing applications (Hendlich et al., 2003; Weisel et al., 2012). Firstly, its very flexible query system allows a user to search for arbitrary geometrical patterns. A query consists of search points, representing atoms in the protein–ligand interface, including atoms from water molecules and metals. The geometrical arrangement can be defined using distance and angle constraints as well as constraints for precalculated interactions (cf. Fig. 3). The molecular environment of a search point can be defined using the powerful SMARTS language. This geometrical search can be combined with numerical and textual constraints for different properties of protein–ligand complexes. Secondly, the mined set of protein–ligand structures is variable and can be freely defined by the user. PELIKAN works with an SQLite database storing all relevant information, which does not need any server infrastructure. Thus, PELIKAN can be used to find specific interaction patterns either within the complete PDB or in all resulting structures of a virtual screening or an MD simulation. On the tools website (http://www.zbh.uni-hamburg.de/pelikan), precalculated databases constructed from different sets of protein–ligand databases are provided.

5. Cheminformatics desktop applications

Processing large molecular data sets is probably the most elementary task in many cheminformatics related areas. Projects related to statistical analyses of chemical libraries, database maintenance, or candidate collection across multiple input sources require stable and efficient software solutions addressing common problems like duplicate removal or calculation of and filtering by molecular properties. Therefore, a consistent handling of molecules from different input formats is an essential basis of every cheminformatics library. Enriching large collections of small organic molecules by additional representations is a fundamental preprocessing step in the workflows of high throughput screening or molecular docking. This mostly requires the enumeration of alternative conformers, tautomers, and protonation states in order to cover a molecule's relevant state space as well as possible. There are many applications which are able to process large amounts of input data, but unfortunately in many cases not all standard file formats are supported. Even more important, file format conversion often comes along with the introduction of erroneous molecule representations. Our molecule converter UNICON (Sommer et al., 2016) bundles the key elements of the NAOMI software library including its consistent chemical model that allows for a stable and highly accurate conversion of the most commonly used file formats like SDF, MOL2,



Status: Found 4 matches in 3 different PDB files (0.0s)

Your current filter chain contains:
0 Standard filters
4 Interaction point filters
3 Point-point constraints
0 Angle constraints

1. Add point filter
- Id: 1 Reference Ligand - Any Type
- Id: 2 Reference Ligand - Any Type
- Id: 11 Protein - Any Type
- Id: 14 Protein - Any Type

2. Add point-point constraint
- Id: 12 First point id: 2 Second point id: 11 Use distance: ☑ min: 2.5 max: 3.5 A Use Interaction type: □
- Id: 13 First point id: 1 Second point id: 11 Use distance: ☑ min: 2.5 max: 3.5 A Use Interaction type: □
- Id: 18 First point id: 1 Second point id: 14 Use distance: ☑ min: 4.5 max: 5.5 A Use Interaction type: □

3. Add angle constraint

Status: Found 4 matches in 3 different PDB files (0.0s)

Your current filter chain contains:
0 Standard filters
4 Interaction point filters
3 Point-point constraints
0 Angle constraints

Sort results by: ☑ EC number □ Molecule

Resulting PDB files:
- Hydrolase: 2O8H, 4DFF
- Lyase: 3PO6

Pockets in 4DFF:
- 0JP_B_801: Match 1

The following pockets are displayed:
- EC Class: Hydrolase PDB ID: 2O8H Ligand: 227_A_999 Match ID: 1 ☑ show
- EC Class: Hydrolase PDB ID: 4DFF Ligand: 0JP_B_801 Match ID: 1 ☑ show

Draw:
□ interactions
☑ results
☑ only relevant residues
□ only rel. ligand atoms
□ sec. struct.

Loaded database:
/home/inhester/calc/interactionDB/TestDatabases/allCompare.sqlite

Fig. 3. Geometrical arrangements of atoms within protein–ligand interfaces can be searched with PELIKAN. Background: Screenshot of the query design tab of PELIKAN. A query consists of a set of search points which match atoms (displayed as green spheres). The spatial arrangement of search points can be defined using distance and angle constraints. The query can be designed from scratch or from a pocket of interest. Foreground: Screenshot of the result presentation tab of PELIKAN. All resulting hits can be visualized in a 3D viewer and superimposed based on the atoms matching the search points from the query. These atoms are highlighted with colored spheres. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

SMILES (Weininger, 1988), InChI (Heller et al., 2013), PDB, and PDBx/mmCIF. The NAOMI molecule representation based on valence states (Urbaczek et al., 2014) also allows us to identify different molecule representations (tautomers and protonation states) as the same molecule and to enumerate alternative molecular states. It is also able to extend the available molecule information with 2D and 3D structure coordinates and generate conformer ensembles. With this comprehensive functionality, UNICON is perfectly suited for all kinds of molecule conversion tasks as well as automated preprocessing of large compound libraries and could easily be integrated into automatic workflows.

Pipeline tools are commonly applied for the processing of chemical datasets as they have the advantage that their functionality is fully automatable and usually also extensible by custom software solutions. On the other hand, they are less flexible if the composition of the data processing workflow depends on intermediate results and requires manual inspection. For this reason, we also offer a solution for the latter scenario. MONA (Hilbig et al., 2013; Hilbig and Rarey, 2015) supports interactive, case-driven preparation and visualization of small-molecule datasets and aims especially at applications where the workflow operations are not known beforehand. MONA integrates solutions for the intuitive visualization of molecular dataset, statistical analyses, filtering properties, and various combinatorial operations for processing multiple molecule sets. Also based on the NAOMI library, MONA can consistently handle molecular entities from different input formats and calculate a broad range of commonly applied molecular properties. The internal application of a molecular database supports instant access to precalculated data, reduces the runtime for many commonly required tasks and thus allows interactive processing of molecule sets with up to a million entries. Furthermore, MONA's graphical interface inspires an intuitive and spontaneous workflow creation and is therefore particularly valuable for users inexperienced with the set-up of more complicated workflow engines or direct database queries.

Besides other criteria, MONA can create molecule sets on the basis of SMARTS. SMARTS patterns are an essential and powerful means for tasks like the description of chemical moieties, database searches, or filtering out undesired molecules from a data set. However, the strong expressive power and the manifold application possibilities of the SMARTS language come along with the disadvantages that SMARTS strings are not easily interpretable and even more difficult to create. With SMARTSeditor (Schomburg et al., 2013), we provide a unique tool that allows the user a convenient and interactive design of complex SMARTS patterns. Based on the intuitive SMARTSviewer (Schomburg et al., 2010) visualization concept, which is derived from structure diagrams, even very complex SMARTS patterns can be depicted in an easily interpretable manner. With the SMARTSeditor functionality, these graphical SMARTS representation can be manipulated interactively. In the case that the user has a certain pattern in mind, a typical SMARTS generation workflow could start with composing a scaffold from the provided set of ring templates and common functional groups by drag-and-drop operations. Afterwards, further elements can be added, edited, or removed interactively. The user can additionally provide a molecule data set, which is visualized in a structure diagram panel and all substructures matching the generated SMARTS pattern will instantaneously be highlighted upon every editing step. Moreover, SMARTSeditor also supports scenarios, where no clear idea of a pattern exist. The integrated SMARTSminer (Bietz et al., 2015) approach allows the user to find frequently occurring patterns in a set of molecules or discriminative patterns that separates one dataset from another. In both cases, the search is not limited to simple substructures but can also derive more generalizing and specifying atomic features. A combination of automatic pattern generation and subsequent interactive refinement constitutes a powerful approach for an efficient design of chemical patterns.

Another cheminformatics-based application is the systematic generation of new molecules, which, e.g., plays an important role in drug design. Several studies have found that known drug molecules can be described by simple physicochemical properties (Lipinski et al., 2001; Ghose et al., 1999; Veber et al., 2002; Reichel, 2006; Oprea et al., 2001). Therefore, molecules with these properties have a high likelihood of being pharmaceutically relevant. In addition, newly designed molecules should be synthesizable. FSees (fragment space exhaustive enumeration system) is an efficient and deterministic method to systematically generate all molecules with a user-defined physicochemical profile (Lauck and Rarey, 2016). The basis for this algorithm are Fragment Spaces (Rarey and Stahl, 2001), a combinatorial chemical space constituting of molecular fragments and connection rules. The latter determine which fragments can be connected and are derived from chemical reactions. In order to apply FSees, a fragment space must be constructed from a library of known molecules (Degen et al., 2008; Lauck and Rarey, 2016). Then this space is enumerated with specific physicochemical constraints thus yielding a library of new molecules. FSees has been applied to different use-cases. First, molecules were constructed from known inhibitors for a specific target. It was shown that the resulting libraries represent a source of promising lead structures. Next, FSees was applied in a fragment-based design context. In this scenario, all generated molecules share a user-defined structural entity, e.g. a certain scaffold. Finally, a library of 0.5 billion lead-like molecules was generated from approved drugs containing mostly novel compounds (Lauck and Rarey, 2016). This library is available for download free of charge from http://www.zbh.uni-hamburg.de/hells.

In order to identify potential drugs for a certain target protein in such large libraries, efficient screening engines are needed. Ligand-based virtual screening can assist addressing this purpose by searching for promising candidates in databases which are too large to be evaluated experimentally. mRAISE (von Behren et al., 2016; von Behren and Rarey, 2017) is a new method which tackles this problem utilizing the previously described screening technology of TrixP and iRAISE (see Section 3) with adaptations to better fit the ligand-based context. Structural ligand alignments calculated based on descriptor matches are scored using atom-centered Gaussian functions in combination with weights representing biochemical similarity or dissimilarity of the respective atoms. The performance of mRAISE is comparable to state-of-the-art methods (Grant et al., 1996; Taminau et al., 2008; Roy and Skolnick, 2015) and due to the index-based search technology used for descriptor comparison, virtual screening with mRAISE on preprocessed databases shows an excellent balance between computing time and result quality compared to other methods.

6. Summary

Computational approaches play an important role in modern life science research. We offer a wide range of software applications that target multiple important issues in the fields related to structural biology, chemistry, medicine, and biotechnology. Our software collection includes many state-of-the-art technologies that are made available either via our unified web service platform Proteins*Plus* or our desktop application package. Due to the increasing amount of biological data, the need for tools to manage and exploit this wealth of information will grow further. While today mostly specially trained experts deal with bioinformatics, we strongly believe that computational methods will belong to the standard repertoire of research tools for every life scientist soon. The Proteins*Plus* server aims at making tools available to non-experts in structural biology. As a future advancement we intend to interconnect various functionalities of Proteins*Plus* in order to provide interactive workflows. With this service and several easy-to-use tools we hope to contribute in paving the road into a digital life science era.

Acknowledgements

The development of Proteins*Plus* is partially funded by de.NBI and the German Federal Ministry of Education and Research (Grant no. 031L0105).

References

An, J., Totrov, M., Abagyan, R., 2005. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol. Cell. Proteomics 4 (6), 752–761.

Arzt, S., Beteva, A., Cipriani, F., Delageniere, S., Felisaz, F., Förstner, G., Gordon, E., Launer, L., Lavault, B., Leonard, G., et al., 2005. Automation of macromolecular crystallography beamlines. Prog. Biophys. Mol. Biol. 89 (2), 124–152.

Bass, M.B., Hopkins, D.F., Jaquysh, W.A.N., Ornstein, R.L., 1992. A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. Proteins: Struct. Funct. Bioinf. 12 (3), 266–277.

Bayden, A.S., Fornabaio, M., Scarsdale, J.N., Kellogg, G.E., 2009. Web application for studying the free energy of binding and protonation states of protein–ligand complexes based on HINT. J. Comput.-Aided Mol. Des. 23 (9), 621–632.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235–242.

Bernauer, J., Bahadur, R.P., Rodier, F., Janin, J., Poupon, A., 2008. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein–protein interactions. Bioinformatics 24 (5), 652–658.

Bietz, S., Rarey, M., 2015. ASCONA: rapid detection and alignment of protein binding site conformations. J. Chem. Inf. Model. 55 (8), 1747–1756.

Bietz, S., Rarey, M., 2016. SIENA: efficient compilation of selective protein binding site ensembles. J. Chem. Inf. Model. 56 (1), 248–259.

Bietz, S., Schomburg, K.T., Hilbig, M., Rarey, M., 2015. Discriminative chemical patterns: automatic and interactive design. J. Chem. Inf. Model. 55 (8), 1535–1546.

Bietz, S., Urbaczek, S., Schulz, B., Rarey, M., 2014. Protoss: a holistic approach to predict tautomers and protonation states in protein–ligand complexes. J. Cheminf. 6 (1), 12.

Block, P., Paern, J., Hüllermeier, E., Sanschagrin, P., Sotriffer, C.A., Klebe, G., 2006. Physicochemical descriptors to discriminate protein–protein interactions in permanent and transient complexes selected by means of machine learning algorithms. Proteins: Struct. Funct. Gen. 65 (3), 607–622.

Brünger, A.T., Karplus, M., 1988. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. Proteins: Struct. Funct. Bioinf. 4 (2), 148–156.

Caboche, S., 2013. LeView: automatic and interactive generation of 2D diagrams for biomacromolecule/ligand interactions. J. Cheminf. 5 (1), 40.

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L.J., Bork, P., 2008. Drug target identification using side-effect similarity. Science 321 (5886), 263–266.

Clark, A.M., Labute, P., 2007. 2D depiction of protein–ligand complexes. J. Chem. Inf. Model. 47 (5), 1933–1944.

Da Silva, F., Desaphy, J., Bret, G., Rognan, D., 2015. IChemPIC: a random forest classifier of biological and crystallographic protein–protein interfaces. J. Chem. Inf. Model. 55 (9), 2005–2014.

Davis, A., St-Gallay, S., Kleywegt, G., 2008. Limitations and lessons in the use of X-ray structural information in drug design. Drug Discov. Today 13 (19–20), 831–841.

Davis, A., Teague, S., Kleywegt, G., 2003. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. Angew. Chem. Int. Ed. 42 (24), 2718–2736.

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., Rarey, M., 2008. On the art of compiling and using 'drug-like' chemical fragment spaces. ChemMedChem 3 (10), 1503–1507.

Fährrolfes, R., Bietz, S., Flachsenberg, F., Meyder, A., Nittinger, E., Otto, T., Volkamer, A., Rarey, M., 2017. ProteinsPlus: a web portal for structure analysis of macromolecules. Nucleic Acids Res. http://dx.doi.org/10.1093/nar/gkx333.

Forneris, F., Heuts, D.P., Delvecchio, M., Rovida, S., Fraaije, M.W., Mattevi, A., 2008. Structural analysis of the catalytic mechanism and stereoselectivity in streptomyces coelicolor alditol oxidase. Biochemistry 47 (3), 978–985.

Gasteiger, J., Engel, T., 2006. Chemoinformatics: A Textbook. John Wiley & Sons.

Ghose, A.K., Viswanadhan, V.N., Wendoloski, J.J., 1999. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J. Chem. Doc. 1 (1), 55–68.

Goldsmith-Fischman, S., Honig, B., 2003. Structural genomics: computational methods for structure analysis. Protein Sci. 12 (9), 1813–1821.

Grant, J., Gallardo, M., Pickup, B., 1996. A fast method of molecular shape comparison: a simple application of a gaussian description of molecular shape. J. Comput. Chem. 17 (14), 1653–1666.

Halgren, T.A., 2009. Identifying and characterizing binding sites and assessing druggability. J. Chem. Inf. Model. 49 (2), 377–389.

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., Pletnev, I., 2013. InChI – the worldwide chemical structure identifier standard. J. Cheminf. 5 (1), 7.

Hendlich, M., Bergner, A., Gnther, J., Klebe, G., 2003. Relibase: design and development of a database for comprehensive analysis of protein–ligand interactions. J. Mol. Biol. 326 (2), 607–620.

Henrick, K., Thornton, J.M., 1998. PQS: a protein quaternary structure file server. Trends Biochem. Sci. 23 (9), 358–361.

Hilbig, M., Rarey, M., 2015. MONA 2: a light cheminformatics platform for interactive compound library processing. J. Chem. Inf. Model. 55 (10), 2071–2078.

Hilbig, M., Urbaczek, S., Groth, I., Heuser, S., Rarey, M., 2013. MONA-interactive manipulation of molecule collections. J. Cheminf. 5 (1), 38.

Hooft, R.W., Sander, C., Vriend, G., 1996. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins: Struct. Funct. Bioinf. 26 (4), 363–376.

Hopkins, A.L., 2008. Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. 4 (11), 682–690.

Inhester, T., Bietz, S., Hilbig, M., Schmidt, R., Rarey, M., 2017. Index-based searching of interaction patterns in large collections of protein–ligand interfaces. J. Chem. Inf. Model. 57 (2), 148–158.

Ivanov, A.A., Khuri, F.R., Fu, H., 2013. Targeting protein–protein interactions as an anticancer strategy. Trends Pharm. Sci. 34 (7), 393–400.

Jalencas, X., Mestres, J., 2013. Identification of similar binding sites to detect distant polypharmacology. Mol. Inf. 32 (11-12), 976–990.

Jones, T., Kjeldgaard, M., 1997. Electron density map interpretation. Methods Enzymol. 277, 173–208.

Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., Shoichet, B.K., 2007. Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. 25 (2), 197–206.

Kellenberger, E., Schalon, C., Rognan, D., 2008. How to measure the similarity between protein ligand-binding sites? Curr. Comput.-Aided Drug Des. 4 (3), 209.

Kleywegt, G.J., 2000. Validation of protein crystal structures. Acta Crystallogr. Sect. D 56 (3), 249–265.

Koeppen, H., Kriegl, J., Lessel, U., Tautermann, C.S., Wellenzohn, B., 2011. Virtual Screening. Wiley-VCH Verlag GmbH & Co. KGaApp. 61–85 Ch. Ligand-Based Virtual Screening.

Krieger, E., Dunbrack Jr., R.L., Hooft, R.W., Krieger, B., 2012. Assignment of protonation states in proteins and ligands: combining pKa prediction with hydrogen bonding network optimization. In: Baron, R. (Ed.), Computational Drug Discovery and Design. Springer, New York, NY, USA, pp. 405–421.

Krissinel, E., Henrick, K., 2007. Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. 372 (3), 774–797.

Kufareva, I., Ilatovskiy, A.V., Abagyan, R., 2012. Pocketome: an encyclopedia of small-molecule binding sites in 4D. Nucleic Acids Res. 40 (D1), D535–D540.

Labute, P., 2009. Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. Proteins: Struct. Funct. Bioinf. 75 (1), 187–205.

Laskowski, R.A., Swindells, M.B., 2011. LigPlot+: multiple ligand–protein interaction diagrams for drug discovery. J. Chem. Inf. Model. 51 (10), 2778–2786.

Lauck, F., Rarey, M., 2016. FSees: customized enumeration of chemical subspaces with limited main memory consumption. J. Chem. Inf. Model. 56 (9), 1641–1653.

Le Guilloux, V., Schmidtke, P., Tuffery, P., 2009. Fpocket: an open source platform for ligand pocket detection. BMC Bioinf. 10 (1), 168.

Leach, A.R., 2001. Molecular Modelling: Principles and Applications. Pearson Education.

Li, X., Jacobson, M.P., Zhu, K., Zhao, S., Friesner, R.A., 2007. Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling. Proteins: Struct. Funct. Bioinf. 66 (4), 824–837.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Rev. 46 (1–3), 3–26.

Lippert, T., Rarey, M., 2009. Fast automated placement of polar hydrogen atoms in protein–ligand complexes. J. Cheminf. 1 (1), 13.

Liu, Q., Li, J., 2010. Propensity vectors of low-ASA residue pairs in the distinction of protein interactions. Proteins: Struct. Funct. Bioinf. 78 (3), 589–602.

McDonald, I.K., Thornton, J.M., 1994. Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. 238 (5), 777–793.

McDonald, I.K., Thornton, J.M., 1995. The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. Protein Eng. 8 (3), 217–224.

Mitra, P., Pal, D., 2011. Combining Bayes classification and point group symmetry under Boolean framework for enhanced protein quaternary structure inference. Structure 19 (3), 304–312.

Nisius, B., Sha, F., Gohlke, H., 2012. Structure-based computational analysis of protein binding sites for function and druggability prediction. J. Biotechnol. 159 (3), 123–134.

Nittinger, E., Schneider, N., Lange, G., Rarey, M., 2015. Evidence of water molecules – a statistical evaluation of water molecules based on electron density. J. Chem. Inf. Model. 55 (4), 771–783.

Oprea, T.I., Davis, A.M., Teague, S.J., Leeson, P.D., 2001. Is there a difference between leads and drugs? A historical perspective. J. Chem. Inf. Model. 41 (5), 1308–1315.

Pawson, T., Nash, P., 2003. Assembly of cell regulatory systems through protein interaction domains. Science 300 (5618), 445–452.

Rarey, M., Stahl, M., 2001. Similarity searching in large combinatorial chemistry spaces. J. Comput.-Aided Mol. Des. 15 (6), 497–520.

Reichel, A., 2006. The role of blood–brain barrier studies in the pharmaceutical industry. Curr. Drug Metab. 7 (2), 183–203.

Rognan, D., 2010. Structure-based approaches to target fishing and ligand profiling. Mol. Inf. 29 (3), 176–187.

Rose, A.S., Hildebrand, P.W., 2015. NGL Viewer: a web application for molecular visualization. Nucleic Acids Res. 43 (W1), W576–W579.

Roy, A., Skolnick, J., 2015. LIGSIFT: an open-source tool for ligand structural alignment and virtual screening. Bioinformatics 31 (4), 539–544.

Schärer, M.A., Grütter, M.G., Capitani, G., 2010. CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. Proteins: Struct. Funct. Bioinf. 78 (12), 2707–2713.

Schlosser, J., Rarey, M., 2009. Beyond the virtual screening paradigm: structure-based searching for new lead compounds. J. Chem. Inf. Model. 49 (4), 800–809.

Schneider, G., Fechner, U., 2005. Computer-based de novo design of drug-like molecules. Nat. Rev. Drug Discov. 4 (8), 649–663.

Schneider, N., Lange, G., Hindle, S., Klein, R., Rarey, M., 2013. A consistent description of HYdrogen bond and DEhydration energies in protein–ligand complexes: methods behind the HYDE scoring function. J. Comput.-Aided Mol. Des. 27 (1), 15–29.

Schneider, N., Volkamer, A., Nittinger, E., Rarey, M., 2016. Applied Biocatalysis: From Fundamental Science to Industrial Applications. Wiley-VCH Verlag GmbH & Co. KGaApp. 71–100 Ch. Supporting Biocatalysis Research with Structural

Bioinformatics.

Schomburg, K., Ehrlich, H.-C., Stierand, K., Rarey, M., 2010. From structure diagrams to visual chemical patterns. J. Chem. Inf. Model. 50 (9), 1529–1535.

Schomburg, K.T., Bietz, S., Briem, H., Henzler, A.M., Urbaczek, S., Rarey, M., 2014. Facing the challenges of structure-based target prediction by inverse virtual screening. J. Chem. Inf. Model. 54 (6), 1676–1686.

Schomburg, K.T., Wetzer, L., Rarey, M., 2013. Interactive design of generic chemical patterns. Drug Discov. Today 18 (13), 651–658.

Sommer, K., Friedrich, N.O., Bietz, S., Hilbig, M., Inhester, T., Rarey, M., 2016. UNICON: a powerful and easy-to-use compound library converter. J. Chem. Inf. Model. 56 (6), 1105–1111.

Stierand, K., Maaß, P.C., Rarey, M., 2006. Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. Bioinformatics 22 (14), 1710–1716.

Stierand, K., Rarey, M., 2007. From modeling to medicinal chemistry: automatic generation of two-dimensional complex diagrams. ChemMedChem 2 (6), 853–860.

Stierand, K., Rarey, M., 2010. Drawing the PDB: protein–ligand complexes in two dimensions. ACS Med. Chem. Lett. 1 (9), 540–545.

Taminau, J., Thijs, G., De Winter, H., 2008. Pharao: pharmacophore alignment and optimization. J. Mol. Graph. Model. 27 (2), 161–169.

Tickle, I.J., 2012. Statistical quality indicators for electron-density maps. Acta Crystallogr. Sect. D 68 (4), 454–467.

Urbaczek, S., Kolodzik, A., Fischer, J.R., Lippert, T., Heuser, S., Groth, I., Schulz-Gasch, T., Rarey, M., 2011. NAOMI: on the almost trivial task of reading molecules from different file formats. J. Chem. Inf. Model. 51 (12), 3199–3207.

Urbaczek, S., Kolodzik, A., Groth, I., Heuser, S., Rarey, M., 2013. Reading PDB: perception of molecules from 3D atomic coordinates. J. Chem. Inf. Model. 53 (1), 76–87.

Urbaczek, S., Kolodzik, A., Rarey, M., 2014. The valence state combination model: a generic framework for handling tautomers and protonation states. J. Chem. Inf. Model. 54 (3), 756–766.

Veber, D.F., Johnson, S.R., Cheng, H.-Y., Smith, B.R., Ward, K.W., Kopple, K.D., 2002. Molecular properties that influence the oral bioavailability of drug candidates. J. Med. Chem. 45 (12), 2615–2623.

Verdonk, M.L., Mortenson, P.N., Hall, R.J., Hartshorn, M.J., Murray, C.W., 2008. Protein–ligand docking against non-native protein conformers. J. Chem. Inf. Model. 48 (11), 2214–2225.

Villoutreix, B.O., Kuenemann, M.A., Poyet, J.L., Bruzzoni-Giovanelli, H., Labbé, C., Lagorce, D., Sperandio, O., Miteva, M.A., 2014. Drug-like protein–protein interaction modulators: challenges and opportunities for drug discovery and chemical biology. Mol. Inf. 33 (6-7), 414–437.

Volkamer, A., Griewel, A., Grombacher, T., Rarey, M., 2010. Analyzing the topology of active sites: on the prediction of pockets and subpockets. J. Chem. Inf. Model. 50 (11), 2041–2052.

Volkamer, A., Kuhn, D., Grombacher, T., Rippmann, F., Rarey, M., 2012. Combining global and local measures for structure-based druggability predictions. J. Chem. Inf. Model. 52 (2), 360–372.

von Behren, M.M., Bietz, S., Nittinger, E., Rarey, M., 2016. mRAISE: an alternative algorithmic approach to ligand-based virtual screening. J. Comput.-Aided Mol. Des. 30 (8), 583–594.

von Behren, M.M., Rarey, M., 2017. Ligand-based virtual screening under partial shape constraints. J. Comput.-Aided Mol. Des. 31 (4), 335–347.

von Behren, M.M., Volkamer, A., Henzler, A.M., Schomburg, K.T., Urbaczek, S., Rarey, M., 2013. Fast protein binding site comparison via an index-based screening technology. J. Chem. Inf. Model. 53 (2), 411–422.

Wang, W., Zhou, X., He, W., Fan, Y., Chen, Y., Chen, X., 2012. The interprotein scoring noises in glide docking scores. Proteins: Struct. Funct. Bioinf. 80 (1), 169–183.

Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28 (1), 31–36.

Weisel, M., Bitter, H.-M., Diederich, F., So, W.V., Kondru, R., 2012. PROLIX: rapid mining of protein–ligand interactions in large crystal structure databases. J. Chem. Inf. Model. 52 (6), 1450–1461.

Wells, J.A., McClendon, C.L., 2007. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. Nature 450 (7172), 1001–1009.

Word, J.M., Lovell, S.C., Richardson, J.S., Richardson, D.C., 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. 285 (4), 1735–1747.

Zhu, H., Domingues, F.S., Sommer, I., Lengauer, T., 2006. NOXclass: prediction of protein–protein interaction types. BMC Bioinf. 7 (1), 27.

Prediction of protein mutation effects based on dehydration and hydrogen bonding – A large-scale study.

[D8] Schomburg, K. T.; **Nittinger, E.**; Meyder, A.; Bietz, S.; Lange, G.; Klein, R.; Rarey, M. Prediction of protein mutation effects based on dehydration and hydrogen bonding – A large-scale study. Proteins Struct. Funct. Bioinforma. 2017, 85 (8): 1550-1566.

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Prediction of protein mutation effects based on dehydration and hydrogen bonding – A large-scale study

Karen T. Schomburg,[1] Eva Nittinger,[1] Agnes Meyder,[1] Stefan Bietz,[1] Nadine Schneider,[1] Gudrun Lange,[2] Robert Klein,[2] and Matthias Rarey [1]*

[1] Universität Hamburg, ZBH - Center for Bioinformatics, Bundestrasse 43, Hamburg, 20146, Germany

[2] Bayer CropScience AG, Industriepark Hoechst, G836, Frankfurt am Main, 65926, Germany

ABSTRACT

Reliable computational prediction of protein side chain conformations and the energetic impact of amino acid mutations are the key aspects for the optimization of biotechnologically relevant enzymatic reactions using structure-based design. By improving the protein stability, higher yields can be achieved. In addition, tuning the substrate selectivity of an enzymatic reaction by directed mutagenesis can lead to higher turnover rates. This work presents a novel approach to predict the conformation of a side chain mutation along with the energetic effect on the protein structure. The HYDE scoring concept applied here describes the molecular interactions primarily by evaluating the effect of dehydration and hydrogen bonding on molecular structures in aqueous solution. Here, we evaluate its capability of side-chain conformation prediction in classic remutation experiments. Furthermore, we present a new data set for evaluating "cross-mutations," a new experiment that resembles real-world application scenarios more closely. This data set consists of protein pairs with up to five point mutations. Thus, structural changes are attributed to point mutations only. In the cross-mutation experiment, the original protein structure is mutated with the aim to predict the structure of the side chain as in the paired mutated structure. The comparison of side chain conformation prediction ("remutation") showed that the performance of HYDE$_{\text{protein}}$ is qualitatively comparable to state-of-the art methods. The ability of HYDE$_{\text{protein}}$ to predict the energetic effect of a mutation is evaluated in the third experiment. Herein, the effect on protein stability is predicted correctly in 70% of the evaluated cases.

Proteins 2017; 85:1550–1566.

Key words: protein side chain conformation; protein stability; mutation effect; bioengineering; protein design; HYDE scoring; dehydration; hydrogen bonding.

INTRODUCTION

Protein mutations have different effects on a protein's function and stability. A mutation can stabilize the protein or tip the complex balance of destabilizing and stabilizing forces within a protein to complete unfolding. Next to the effects on protein stability, mutations can also have large effects on the protein's function. Exchanging an amino acid certainly has numerous effects, for example, altering the proteins enzymatic function, its substrate specificity, or even blocking its function completely.[1]

In pharmaceutical sciences, it is often crucial to understand the effect of a natural mutation on a drug target protein. Is the modified protein still binding the drug compound? What changes in specificity and binding free energy are related to the mutation? In biotechnology, a protein's function is exploited to get a profitable product. Often much better yields can be achieved if a protein's function can be enhanced by a favorable mutation. Such a mutation might improve the thermostability of the protein and therefore allows running the reaction at higher temperatures and prolonging the lifetime of the

Additional Supporting Information may be found in the online version of this article.
This work was performed at Universität Hamburg, ZBH - Center for Bioinformatics, Bundestrasse 43, 20146 Hamburg, Germany.
Grant sponsor: BMBF; Grant number: 031A183B.
K. T. Schomburg's present address is OSTHUS GmbH, Eisenbahnweg 9-11, Aachen, 52068, Germany
*Correspondence to: Matthias Rarey; Universität Hamburg, ZBH - Center for Bioinformatics, Bundestrasse 43, 20146 Hamburg, Germany. E-mail: rarey@zbh.uni-hamburg.de
Received 11 January 2017; Revised 29 April 2017; Accepted 2 May 2017
Published online 9 May 2017 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.25315

protein in reaction conditions. Furthermore, a mutation might shift or broaden the substrate specificity of an enzyme.

One possibility to explore mutations is random or directed experimental mutagenesis.[2–4] During the past decades, however, more rational computer-based approached have been pursued. Herein, the Rosetta[5] software package has a very broad impact among other existing methods.[6] Two recent examples are the increase of the thermostability of an *Aspergillus oryzae* cutinase with rational design applying Rosetta modeling[7] and the engineering of a transcription factor with the Rosetta protocol.[8] Using these rational computational approaches, the number of experiments can be reduced substantially.

Recent methods are frequently based on exploiting the high amount of available data to derive computational models. A basis for such an approach can be common amino acid sequences or amino acid conformations often observed in crystal structures. While this approach is certainly a sensible one, the developed model is often highly complex and does not give rational insights. The question we are addressing in this manuscript is whether we are already at the point of predicting the energetic impact of mutations on a proteins structure based on modeling of interactions on the atomic level using a physics-based approach.

There are two main questions which could be addressed by computational mutation prediction. (1) Which effect does a mutation have on the protein stability? (Right now most of the tools focus on the effect on protein stability and not on the effect on the function, exceptions are reviewed by Henikoff *et al.*[9]) (2) How does the structure of the mutated protein look like (prediction of the mutated side-chain conformation and its close environment)? The answer to the first question is mainly important if amino acids far from the active site or protein interfaces are mutated and the conformation is less relevant. The answer to the second questions is important if steric and physicochemical features within the active site, within a tunnel to the protein's active site, or a protein's interface are crucial to the protein's function.

Available methods can be grouped into those estimating $\Delta\Delta G$ differences (energy difference between the protein's original and mutated amino acid resulting from noncovalent interactions), thus the overall energetic effect on the protein stability, and those focusing on the geometry for predicting the resulting protein structure including the side-chain conformation of the mutated amino acid. The performance of a method predicting $\Delta\Delta G$ values can be assessed by correlation with measured $\Delta\Delta G$ values or, more coarse-grained, by comparing the sign of $\Delta\Delta G$ values (i.e., does the method predict correctly if a mutation is stabilizing or destabilizing). The ability of a method to predict a protein conformation correctly can be measured by comparing RMSD or Chi(X)-angle differences of predicted amino acid conformations. Herein, existing crystal structures of proteins are used to predict the mutated amino acid side-chain conformation while keeping all surrounding amino acids fixed. Generally, two Chi-angles are measured for the mutated amino acid as a success criterion, χ1-angle (N-CA-CB-CG) and χ2-angle (CA-CB-CG-CD). For a successful prediction, both χ-angles have to be within ±40° of the crystal conformation.

The computational methods predicting $\Delta\Delta G$ values are usually divided into sequence- or structure-based methods, which are combining descriptors with machine learning approaches. MuPro is an example of a sequence-based method using a support vector machine.[10] The tool achieves an accuracy of 84% when classifying the change in the $\Delta\Delta$G value. Comparing protein sequence versus tertiary structure as input the methods encounter nearly no performance difference. Structure-based methods can also rely on computational energy-based modeling such as molecular dynamics, Monte Carlo, force-field-based approaches, and statistical energy potentials (see Masso *et al.*[11] for an overview). The most prominent approaches are AUTO-MUTE,[11] CUPSAT,[12,13] Prethermut,[14] Dmutant,[15] FoldX,[16] PoPMuSiC,[17] iMUTANT,[18,19] MuPro,[10] MultiMutate,[20] SCide,[21] SRide,[22] and Scpred.[23]

Khan and Vihinen[24] compared the predictive power of sequence- and structure-based tools on ProTherm[25] data and find varying strength in predictive power. ProTherm is the data set most often used for evaluation and training of mutation prediction methods. The collection of protein mutations from literature, annotated with thermodynamic data, and links to other database like the Protein Data Base is accessible via a webserver.[26] I-Mutant 3.0 achieved the best performance with an accuracy of 0.64. They also point out that they often obtained worse results than the authors, showing that each tool clearly performs best in the hands of the expert. Potapov *et al.*[27] compared energy-function based methods: CC/PBSA,[28] EGAD,[29] FoldX,[16] I-Mutant2.0,[19] Rosetta,[30] and Hunter.[31] They observe that all methods are able to predict a correct trend but fail in the details. In their evaluation, the best correlation coefficients between experimental and predicted $\Delta\Delta G$ values were around 0.59 while the worst performance resulted in an *R*-value of 0.26. Most recent methods such as MAESTRO[32] aim at improving the prediction of the $\Delta\Delta G$ trend with a machine learning approach. On a ProTherm subset, they achieve similar results as others in the field with a Pearson's correlation coefficient of about 0.7 for predicted versus experimental $\Delta\Delta G$ values.

Protein side-chain conformation prediction has already been addressed for a long time.[33] Existing methods can be differentiated according to their main solution strategy into four main groups: (1) knowledge-based rotamer library approaches, (2) energy functions, (3) machine learning, and (4) statistical methods.[34–40] The conformational space is commonly explored using the backbone-dependent rotamer library developed by Dunbrack *et al.*[34–37] The frequency of observed conformations in experimental protein structures is also used to derive a probability applied in pseudoenergy functions. With their latest approach (SDWRL4[37]), Dunbrack *et al.* achieve an accuracy of the $\chi 1$-angle of 86% and $\chi 1 + \chi 2$ of 75% and for high-resolution protein structures 89% for $\chi 1$ and 80% for $\chi 1 + \chi 2$. Recent approaches refine these frequency-based rotamer libraries with other criteria like energetic effects to reduce the needed number of conformers for sampling.[38] Energy-based approaches make use of force fields such as CHARMM to predict side-chain conformations[40] or train a scoring function based on physicochemical terms; that is, contact surface, volume overlap, backbone dependency, electrostatic interactions, or desolvation energy.[39] Both approaches achieve similar accuracies like the rotamer-based approaches, while especially the force-field approaches are computationally more time consuming.

Machine learning methods, knowledge-based derived rotamer libraries and energy functions, and statistical methods all depend on the data available for learning. While the amount and quality of available data continuously grows, these methods will nevertheless always depend on the cases most prominently present in the data and the assumption that most prominent features are indeed energetically preferred. In contrast to these approaches, we are focusing on the ability to predict mutation effects by computationally modeling molecular interactions at the atomic level using simple physics. In the HYDE energy function applied here, molecular interactions are modeled in great detail, trying to map the energetic landscape in geometric models.[41] Predicting mutation effects in proteins can show how well these modeling approaches are in practice today. HYDE$_{\text{protein}}$ allows $\Delta\Delta G$ difference and residue conformation analysis. Together with a numerical optimizer, the optimal conformation is identified and a score assesses the energetic difference to a mutated structure.

In this publication, we introduce the application of the HYDE scoring function to protein mutational effects (HYDE$_{\text{protein}}$) including a freely available, new data set for benchmarking mutation prediction tools. A comparison to conventional methods is of special interest as HYDE$_{\text{protein}}$, based on a generic physical approach, was not developed for mutational prediction. HYDE$_{\text{protein}}$ results are compared with those of Khan *et al.*[24] Furthermore, a variety of experiments are performed to assess the performance of HYDE$_{\text{protein}}$: remutation, cross-mutations, and protein stability predictions. To perform these experiments, we introduce a new data set of about 10,000 pairs of crystal structures containing mutations of which about 9,000 remain when mutations to PRO and ALA are omitted. In this contribution, besides the introduction of the HYDE$_{\text{protein}}$ approach, we want to establish an evaluation strategy more extensive and closer to real-world applications.

METHODS

For predicting side-chain conformations and energetic mutation effects, two steps have to be performed: generation of conformations of the mutated amino acid and ranking of these. Ideally, a conformation highly similar to the one found in a crystal structure is ranked first.

In the following, we describe how we implemented those two steps followed by the three evaluation experiments remutation, cross-mutation, and energy prediction of mutations. Finally, the data sets used for the evaluation experiments are described.

Steps of the computational mutation strategy

Amino acid mutation generation

In a protein structure, a mutation is carried out as follows: First, the side-chain atoms of the original residue are removed from the structure. Then a default configuration of the mutated amino acid is added to the mutation site, with the backbone atoms guiding the transformation of the atom coordinates. Next, conformations are generated for the amino acid, starting from the default conformation by systematic rotation at each rotatable bond. Here, an enumeration approach is chosen, going from fine to coarse granules to avoid conformational explosion for amino acids with flexible chains. The first rotatable bond is rotated in 60° steps, the second in 90° steps, the third in 120° steps, and the fourth one is only rotated once. With this approach, the number of conformations is kept on a computationally manageable level, for example, 72 rotamers are generated for GLU and only six for SER. Note that these conformations are considered as starting points for the subsequent optimization and are therefore not meant to be in energetic local optima.

Optimization of the amino acid conformation

Before scoring the quality of the given conformation, it is necessary to optimize the conformation in accordance to the degrees of freedom of the scoring function. The objective function of the GeoHYDE optimizer applied here consists of the interaction model of HYDE with modifications to make it numerically stable in combination with a Lennard–Jones potential to avoid atomic clashes.[41] In addition, GeoHYDE contains terms to guarantee low-energy conformations for both the considered amino acid and the surrounding protein: all amino acid conformations are evaluated based on a customized intramolecular Lennard–Jones potential in combination with a continuous knowledge-based torsion potential based on the TorsionLib.[42,43] For the optimization, all rotatable bonds including those to hydrogen-bond donors are fully flexible. Note that bond length and other bond angles are not modified. Owing to the complexity of the GeoHYDE energy landscape, the search for the minimum can only take place locally, making multiple conformations as starting points for the optimization necessary. More implementation details of the optimization are given in Section 1 of the Supporting Information.

For most of the experiments, only the mutated residue is flexible while the rest of the protein is kept rigid. However, for the cross-mutation and for the prediction of a mutation effect, we tested the $\text{HYDE}_{\text{protein}}$ performance with a flexible active site (active site residues that have at least one heavy atom within a pocket of 6.5 Å around the mutated residue).

Scoring of amino acids with $\text{HYDE}_{\text{protein}}$

Before scoring with $\text{HYDE}_{\text{protein}}$, a preprocessing of the structures is undertaken: crystallographic water molecules are optionally removed, hydrogens are assigned, and tautomeric and protonation states and hydrogen orientations are calculated with Protoss.[44] $\text{HYDE}_{\text{protein}}$ applies the basic principles of the HYDE scoring function[41] to an amino acid. The HYDE scoring function incorporates three main energetic effects: hydrogen bonding, the hydrophobic effect, and desolvation. In protein–ligand complexes, HYDE estimates the energetic differences between the unbound and the bound state. It is not calibrated on affinity data of small molecules. Instead, log P increments are derived from a set of small molecules with experimentally measured values.[45] Owing to the general concept of the HYDE scoring function and the independence of experimentally measured affinity data, its scoring concepts can also be applied to score protein–protein interactions or energetic contributions of single amino acid side chains. The free energy estimated by HYDE for a single side-chain represents the difference between the unfolded primary protein chain and its folded, tertiary structure.[46]

The amino acid atom scores are summed up into a total HYDE-score. For assessing not only the score of the investigated amino acid itself but also the contribution of the surrounding amino acids, a score is calculated as the sum of all $\text{HYDE}_{\text{protein}}$ scores for all residues in the pocket. The pocket of a residue consists of all other residues that are within a 11.5 Å radius of the mutated residue. This "large" site was chosen to better describe mutations of smaller residues such as GLY into larger ones such as LYS and ARG. If water molecules, ions, or ligands are part of the pocket, their score is also added to the total $\text{HYDE}_{\text{protein}}$ score. For a classification of the amino acid mutation, that is, whether it is energetically favorable or unfavorable, the original amino acid is scored in the same way. In Figure 1, an example for $\text{HYDE}_{\text{protein}}$ scores for an unfavorable [Fig. 1(A)] and a favorable [Fig. 1(B)] residue are shown. The example protein is a β-glucosidase chosen from the data set used for remutation analysis (see the section "Remutation experiment"). In this figure, the intuitive HYDE atom coloring is shown where red-colored atoms being energetically unfavorable and green-colored atoms are energetically favorable, while white-colored atoms are neutral.

Characterization of amino acids

To evaluate and analyze the performance of our method in different scenarios, we classified amino acids as being "buried" or "solvent accessible." For this, we used the rather strict criterion that a buried amino acid must have at least 20 other residues in its proximity, that is, within a radius of 6.5 Å.

Evaluation experiments

Remutation experiment

In a remutation experiment, a single amino acid is removed from the structure and inserted back keeping the rest of the structure constant including crystallographic water in place. The original amino acid conformation is scored and then a side-chain conformation prediction is carried out as described above, independent from the original conformation, and scored as well.

As performance metrics, we used the ones that are prominent in the scientific field: RMSD and Chi (χ)-angle differences. For calculating RMSD, only the protein side-chain atoms without Cβ are used. A closer look reveals that the two metrics can give a substantially different picture. Even a conformation with large χ-angle differences can have a small RMSD, see example in Figure 2. Here, an ARG is remutated and the found conformation has a low RMSD (0.5 Å) but high χ^1 and χ^2 angle differences (43.3° and 67.2°) and thus low rotational accuracy. However, all interactions would be achieved with the predicted conformation.

In the field of predicting amino acid side-chain conformations, mostly a χ-angle cutoff of ±40° is used to classify a prediction as successful. We used this cutoff as well, to allow a comparison to other methods. Additionally, we evaluated a smaller cutoff of 20°, which in our view is more suitable to detect correct predictions. We calculate RMSD and χ-angle differences not only to the original amino acid conformation in the crystal structure but also to the optimized original amino acid conformation. This conformation was optimized the same way as the mutated amino acid conformation.

Cross-mutation experiment

We created a new experiment called "cross-mutation," referring to cross-docking experiments, where a ligand from one co-crystallized complex is docked into a complex which is crystallized with a different ligand. In a

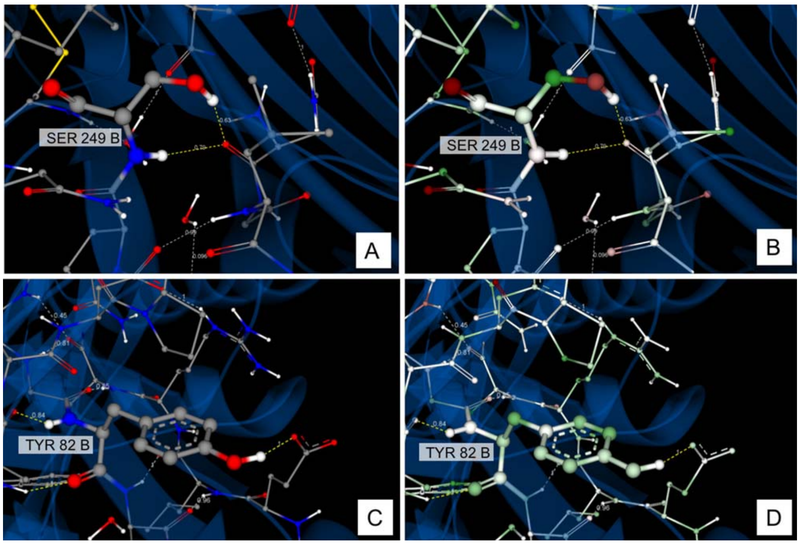


Figure 1

Example of scoring amino acids of a β-glucosidase (PDB code 4I3G). (**A**) SER 249 B in atom colors. (**B**) SER 249 B in HYDE scoring colors. The residue gets an unfavorable HYDE$_{\text{protein}}$ score of 12.52. While the hydrogen donor of the hydroxyl group forms a weak hydrogen bond, the hydrogen bond acceptor cannot engage in any hydrogen bond, thus is highlighted in red. The same applies for the backbone oxygen, which is also highlighted in red. (**C**) TYR 82 B in atom colors. (**D**) TYR 82 B in HYDE scoring colors. The residue gets a favorable HYDE$_{\text{protein}}$ score of −11.45. The hydroxyl group forms a hydrogen bond with ideal geometry and the hydrophobic atoms are in contact with hydrophobic partners.

cross-mutation experiment, we take two crystal protein structures with one (or more) mutation(s) and mutate the residues into the amino acid types of the other structure, then compare the conformation with the crystallized mutation. A schematic example can be found in Figure 3.

A data set of protein pairs containing 1–5 mutations was compiled (see the section “Data sets”). For each single mutation pair (one mutation only), we did the cross-mutation experiment in both directions. For the multiple mutation pairs (more than one mutation), we did each mutation separately. For the cross-mutations, all crystal waters were removed before structure prediction and scoring.

To investigate the mutual interference of several mutations, we analyzed if an additional described mutation occurs in the pocket around the mutated amino acid. As a performance metric, we used RMSDs and χ-angle differences, comparing the conformation of the mutated amino acid to the conformation of the crystallized conformation. For the RMSD calculation, the complete backbones were used for the superposition, which may result in higher values if there is backbone flexibility. The cross-mutation experiment is much more difficult than the remutation experiment, as the surrounding amino acids might have different conformations in the crystal structure. In those cases, for which the mutation with rigid amino acids in the pocket resulted only in clashing conformations, we repeated this experiment with a flexible pocket in which amino acids within the pocket

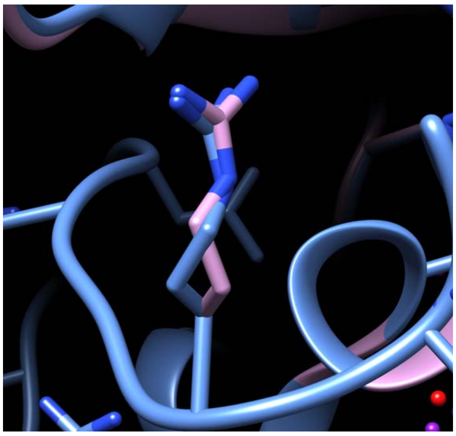


Figure 2

Remutation example where a low RMSD is reached (ARG 345 B, PDB code 1UR1), but the rotational accuracy is low; pink = crystal structure, blue = mutated amino acid conformation. [Color figure can be viewed at wileyonlinelibrary.com]

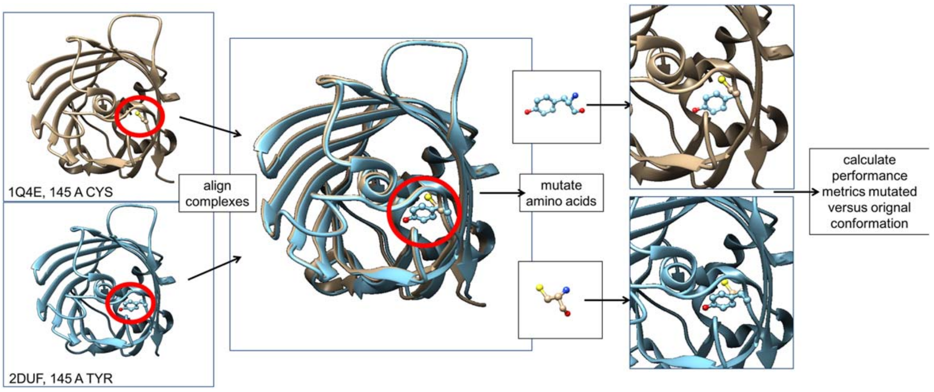

1Q4E, 145 A CYS
2DUF, 145 A TYR
align complexes
mutate amino acids
calculate performance metrics mutated versus orignal conformation

Figure 3

Schematic representation of a cross-mutation experiment. Two complexes containing a mutation are aligned. For each complex, the amino acid is mutated into the one from the other complex. Finally, the conformation of the mutated amino acid is compared to the original crystal conformation in the other complex. [Color figure can be viewed at wileyonlinelibrary.com]

(6.5 Å around the investigated amino acid) are allowed to rotate during the optimization step.

Energy prediction of mutations

In the third experiment, we wanted to assess how well our method is able to predict if a mutation has a stabilizing, a destabilizing, or no effect on the protein stability. We compare the HYDE$_{\text{protein}}$ score of the mutated amino acid to the HYDE$_{\text{protein}}$ score of the original amino acid: if the difference is negative, the mutation is expected to be energetically stable otherwise unstable.

Khan *et al.* evaluated the ability of methods to predict neutral mutations ("negatives") versus mutations of having "any" effect on the protein stability ("positives"). In our evaluation, however, we decided to evaluate if our method predicts the correct effect, that is, stabilizing, destabilizing, or no effect. Thus we applied two evaluation strategies: First, we put the neutral cases aside and evaluated if our method correctly identifies the stabilizing and destabilizing cases simply by evaluating the sign of the HYDE-score difference of the mutation. Second, we tested if our method is able to categorize correctly all three cases. For the first evaluation experiment, we calculated the same performance metrics as Khan *et al.* in their evaluation: accuracy, specificity, sensitivity, and Matthew's correlation coefficient (see Khan *et al.*[24] for definitions). For the second evaluation experiment, we calculated the percentage of correct predictions for all classes separately next to the accuracy.

For this experiment, the mutation was carried out as described in the section "Steps of the computational mutation strategy." The PDB structures were stripped of all crystal waters before the mutation. Default parameters were used otherwise. A flexible pocket was used for mutations leading only to clashing solutions (see the section "Cross-mutation experiment").

Data sets

Remutation data

For the remutation experiment, we created a data set of 100 random crystal structures as no established benchmark data set exists for this experiment. We only selected high-resolution crystal structures from the data set of Nittinger *et al.*[47] and remutated all amino acids except ALA and GLY. ALA and GLY are omitted because they have no alternative conformations. For a representative number of reliable protein side-chain conformations, we randomly selected 100 high-resolution crystal structures with available electron density. We decided not to use the data set compiled by Pottel *et al.*,[48] who used a set of 98 PDB structures, because their set was selected focusing on mutations in the protein–ligand interface. A list of the PDB codes of these 100 crystal structures is provided in Section 2 of the Supporting Information.

Cross-mutation data set

For our new developed cross-mutation experiment, we compiled a new data set. The PDB structures were selected from the same data set of high-quality crystal structures as above (criteria: high resolution ($\leq$1.5 Å) and electron density data available).[47] Within this data set, we searched for pairs of homologous proteins deviating only by 1 (single) to 5 (multiple) mutations with the tool SIENA.[49] SIENA assembles ensembles based on the alignment algorithm ASCONA[50] (for details, see

Table I

Counts of Amino Acid Mutations in the Cross-Mutation Data Set. [Color table can be viewed at wileyonlinelibrary.com]

Target amino	Starting amino acid																	
acid	SER	THR	CYS	VAL	LEU	ILE	MET	PHE	TYR	TRP	ASP	GLU	ASN	GLN	HIS	LYS	ARG	TOTAL
ALA	368	18	178	185	44	424	13	245	7	2	6	59	24	66	1	6	49	1695
GLY	22	34	0	2	0	32	0	0	0	0	98	3	6	4	0	0	21	222
SER		43	63	3	0	0	0	1	1	3	0	2	0	0	0	0	1	117
THR	7		17	39	3	162	0	0	186	0	1	0	0	0	0	0	0	415
CYS	48	15		0	0	0	0	0	27	0	0	0	0	0	0	0	0	90
VAL	34	138	0		325	306	18	10	0	0	24	1	0	0	52	12	0	920
LEU	26	3	1	401		166	185	286	3	3	2	9	24	1	55	2	1	1168
ILE	12	186	0	301	162		25	84	0	4	14	101	62	6	0	0	0	957
MET	0	58	0	22	261	223		0	0	0	0	0	0	304	0	1	25	894
PHE	1	0	4	5	294	86	51		156	19	0	0	0	0	3	0	1	620
TYR	1	162	16	0	2	24	0	76		0	5	0	0	0	0	0	0	286
TRP	0	0	0	0	3	0	0	40	0		0	0	0	0	0	0	1	44
ASP	0	3	0	11	2	14	0	0	0	0		22	244	5	0	0	0	301
GLU	1	0	4	0	6	88	0	0	0	0	16		0	146	9	0	2	272
ASN	6	0	0	0	0	18	0	0	0	0	219	0		3	3	0	0	249
GLN	0	18	0	0	1	6	201	0	0	0	5	147	0		18	0	9	405
HIS	0	0	0	22	57	0	0	4	0	0	31	5	19	23		0	2	163
LYS	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0		6	9
ARG	0	0	0	0	1	0	24	0	0	0	0	2	0	0	0	1		28
SUM	526	678	283	991	1163	1549	518	746	380	31	421	351	379	558	141	22	118	

Supporting Information, Section 3). In this study, we applied SIENA for searching alternative structures of the whole query protein using the following filter criteria: sequence identity parameter = 0.9; maximum number of mutations = 5; and maximum total backbone RMSD = 3 Å. This way, a sequence similarity of 90% and the additional total limit of 5 mutations (replacements) in the calculated global sequence alignment are ensured. The additional RMSD cutoff was applied to eliminate strongly deviating protein conformations.

For each protein, a list of conformations containing the following information is generated: the partner PDB code, the backbone RMSD of the aligned complexes, the position and type of original amino acid, the position and type of mutated amino acid, and the number of mutations between these two proteins. We filtered the data set to contain only standard amino acids (not modified) with valid coordinates for each atom. Furthermore, we excluded structures where either the original or the mutated site contains a ligand; as this could influence the conformation of the amino acid and thus makes them not comparable. Furthermore, only buried amino acids were kept in this experiment. Please note that the terms "mutated" and "original" here are just used for convenience, the assignment of wild type and mutated is not of relevance here.

The data set can be found in the Supporting Information. There the PDB codes for each pair and the amino acid mutations are listed with protein chain identifier, residue position, and amino acid codes. For completeness, we listed in the Supporting Information all protein PDB pairs, which are in total 10,002. For the evaluation, however, we used only those cross-mutations where the target is not ALA or GLY or PRO. This leaves in total 8,855 cases.

Table I gives an overview of the number of mutations within the cross-mutation data set: For each amino acid (row), the count of how many times it was mutated to another amino acid (column) is given. Please note here: first, the distribution is not equal; some mutations are more likely than others, like VAL to LEU and vice versa. Additionally, for many combinations, no data is present as this data set relies on the available PBD structures. Second, not all mutations are present in both directions, for example, the mutation LEU to VAL is counted 401 times, the mutation VAL to LEU only 325 times. The reason for this is that each amino acid is classified as "buried" separately. As we only consider conformation predictions of mutations to buried amino acids, some of the pairs were discarded.

Energy prediction data

For the energetic classification of mutations, we used the data set that Khan and Vihinen[24] compiled for their evaluation of $\Delta\Delta G$ prediction tools. Adopting their evaluation strategy, we defined mutations within a $\Delta\Delta G$ of $\pm$ 0.5 kcal/mol as neutral, $\Delta\Delta G > 0.5$ kcal/mol as stabilizing, and $\Delta\Delta G < -0.5$ kcal/mol as destabilizing (please note the sign of the $\Delta\Delta G$ value: in ProTherm, a positive $\Delta\Delta G$ indicates stabilization and a negative $\Delta\Delta G$ destabilization). All required structures were downloaded from PDB (www.pdb.org) and used without modification.

Some of the mutations Khan *et al.* extracted from ProTherm[25] are omitted because the amino acids are incomplete or modified in the downloaded crystal structure (see Supporting Information, Section 4). In total, the data set

Table II

Results of Conformation Prediction for Flexible Amino Acids With Crystal Water (wW) and Without Crystal Water (woW)

AA	Without/with water	% RMSD < 1.0 Å	Mean RMSD	% X1 < 20° (<40°)	% X1 + X2 < 20° (40°)
SER	woW	61.50	0.91	55.67(60.83)	
	wW	66.83	0.81	61.17(65.83)	
THR	woW	80.87	0.53	74.76(80.55)	
	wW	87.64	0.35	82.02(87.48)	
VAL	woW	87.72	0.30	72.16(74.67)	
	wW	88.49	0.29	72.71(75.37)	
CYS	woW	89.77	0.37	84.66(91.48)	
	wW	89.20	0.39	84.66(90.34)	
ASN	woW	87.34	0.54	86.39(91.77)	72.15(84.49)
	wW	93.33	0.38	93.65(97.78)	73.02(86.98)
ASP	woW	89.58	0.46	87.26(93.82)	74.13(87.64)
	wW	95.38	0.32	95.00(98.85)	76.92(92.69)
HIS	woW	92.38	0.41	94.29(95.71)	78.57(86.19)
	wW	93.30	0.30	95.22(95.69)	79.43(89.95)
ILE	woW	77.28	0.67	80.79(88.19)	61.59(67.90)
	wW	78.11	0.64	81.35(88.74)	63.42(68.83)
LEU	woW	79.53	0.59	69.80(83.64)	58.69(64.06)
	wW	81.17	0.55	71.07(85.09)	60.01(65.32)
PHE	woW	97.11	0.31	97.66(99.86)	**96.69(98.48)**
	wW	97.38	0.30	98.07(100.0)	**96.97(98.76)**
TRP	woW	93.08	0.53	94.62(96.92)	**93.85(93.85)**
	wW	98.46	0.33	98.46(98.46)	**98.46(98.46)**
GLN	woW	64.48	1.06	68.85(78.69)	60.66(67.21)
	wW	85.08	0.54	86.74(91.71)	82.32(85.64)
GLU	woW	81.41	0.69	77.89(87.44)	71.86(82.91)
	wW	87.94	0.45	89.45(93.97)	82.41(90.95)
MET	woW	51.86	0.92	69.34(82.52)	53.87(63.04)
	wW	51.00	0.93	69.34(80.80)	51.29(60.74)
TYR	woW	96.63	0.28	97.62(98.41)	**93.65(97.82)**
	wW	97.02	0.21	97.82(97.82)	**92.86(97.22)**
LYS	woW	48.57	1.35	60.00(73.33)	45.71(53.33)
	wW	51.43	0.97	66.67(82.86)	50.48(62.86)
ARG	woW	60.61	**1.37**	**63.64**(78.79)	55.15(63.64)
	wW	84.85	**0.65**	**80.61**(91.52)	72.73(80.61)

The percentage of conformations below 1.0 Å RMSD, the mean RMSD, the χ^1-angle differences below 40° and 20°, and the $\chi^1 + \chi^2$ differences below 40° and 20° are given. Bold numbers are discussed in detail in the "Results" section.

comprises 1721 data points, of which 182 are stabilizing, 652 neutral, and 887 destabilizing mutations.

Two obstacles have to be faced with the ProTherm data set of Khan *et al.*: first, not all the crystal structures in this data set are of high quality; second, the methods evaluated by Khan *et al.* are all evaluated on different subsets of this data. As some of these methods were trained on parts of the ProTherm data set, Khan *et al.* excluded the data from their evaluation. Therefore, each method is evaluated on a different subset of the whole data set, making a direct comparison impossible. To respect the different structural qualities, we categorized the data into "all cases," "medium-resolution cases" (PDB structures with a resolution <2.5 Å), and "high-resolution cases" (PDB structures with a resolution <1.5 Å).

RESULTS AND DISCUSSION

Remutation results

In this experiment, we remutated all buried residues of 100 random high-resolution PDB complexes into themselves; the original conformation was neglected during the prediction. We compared the best scored conformation according to $\mathrm{HYDE}_{\mathrm{protein}}$ to the original crystal conformation. Two scenarios were considered in the conformation prediction: all crystal waters were removed from the complex and all crystal waters were kept present in the complex. In Table II, the results are summarized: for each amino acid, the percentage of RMSDs smaller than 1.0 Å, the mean RMSD, the percentage of the first χ^1-angle smaller 20° and 40°, and the percentage of the first two χ angles smaller 20° and 40° are listed. We consider all amino acids except ALA, GLY, and PRO; here, the conformational space is not affected by rotational freedom.

Including crystal waters in the conformation prediction improves the performance in almost all cases (better results with waters included for (a) RMSD <1.0 Å: 15/17 amino acids; (b) $\chi1 < 20°$: 15/17 amino acids; (c) $\chi1 + \chi2 < 20°$: 11/13 amino acids). This effect is expected as the rotational freedom of an amino acid is often restricted by water molecules and the conformation

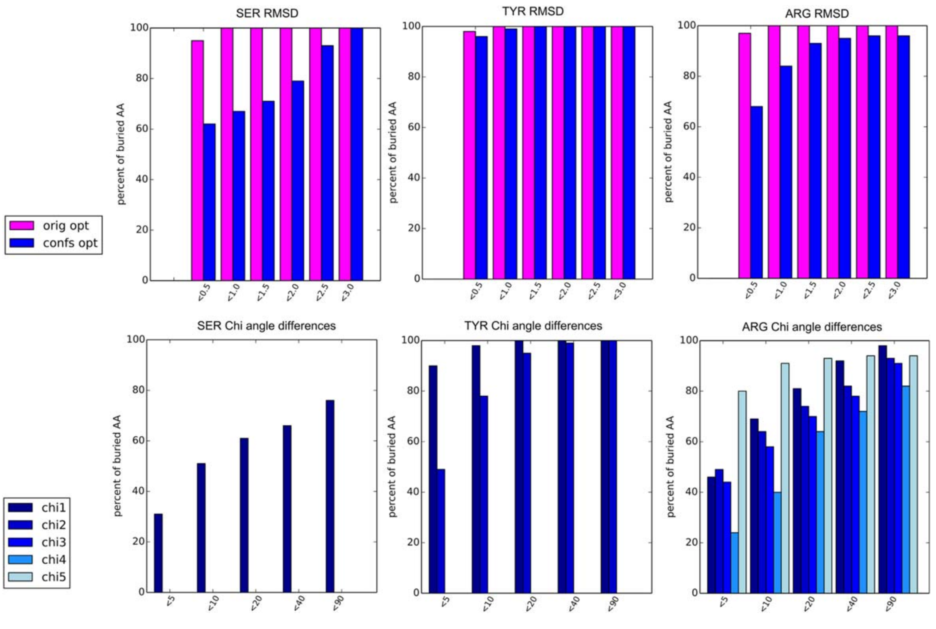


Figure 4

Detailed performance metrics of remutation for SER, TYR, and ARG. Top row: histograms for the RMSDs of SER, TYR, and ARG below 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 Å thresholds. Pink bars: RMSD of optimized crystal conformation versus crystal conformation. Blue bars: remutated conformation versus original conformation. Lower row: histograms for X angle differences below 5°, 10°, 20°, 40°, and 90° cutoffs of remutated versus original conformation, chi1–5: χ-angles starting with $\chi 1$ (N-CA-CB-CG) to $\chi 5$ (CD-NE-CZ-N) for arginine. [Color figure can be viewed at wileyonlinelibrary.com]

might be supported by water-mediated hydrogen bonds. This result is further supported by the fact that side chains with more rotatable bonds have a greater decrease in performance if water molecules are not included in the conformation prediction; that is, ARG achieves χ^1 differences <20° of 81% with water molecules but only 64% without water molecules.

Our method performs best on large aromatic amino acids such as TYR (97% of $\chi^1 + \chi^2 < 40°$), TRP (98% of $\chi^1 + \chi^2 < 40°$), and PHE (98% of $\chi^1 + \chi^2 < 40°$). In the case of ARG, we achieve a satisfying performance with a mean RMSD of 0.65 Å by including the crystal waters. However, ignoring the water molecules, the performance decreases to a mean RMSD of 1.37 Å. For SER, TYR, and ARG, the histograms for the χ^1 and the RMSDs are shown in Figure 4.

As a baseline experiment, we tested if our optimization algorithm is able to identify the crystal conformation of SER, TYR, and ARG. Therefore, we optimized the crystal conformation of these residues. Almost all the optimized conformations were found with an RMSD below 0.5 Å (see Fig. 4, pink bars). In the remutation experiment, the correct conformations (RMSD below 0.5 Å) are almost always found for TYR. For SER and ARG, on the other hand, the correct conformations are only identified in two-thirds of the cases (see Fig. 4, blue bars).

The results show that SER is the most difficult to predict, which was also observed by others in the field.[37] An example is shown in Figure 5: the original conformation of SER 935 A of a clathrin adaptor protein (PBD code 1KYF) is scored with a HYDE-score of 2.4 as energetically unfavorable. In this conformation, the hydrogen acceptor of the hydroxyl group is interacting with the amide nitrogen of ASN-A-852 and the hydrogen donor forms a weak H-bond with the carbonyl-oxygen of CYS-A-931. In the remutation, the best scored conformation has a high RMSD (2.6 Å) and a large χ^1 angle difference (163°) to the original conformation. In this HYDE$_{\text{protein}}$-favored conformation, the hydroxyl group can form multiple interactions with a water-cluster nearby resulting in an energetically favorable score of −0.3.

Table III summarizes the mean and median RMSD and the percentages of χ^1 and $\chi^1 + \chi^2$ smaller than 40° and 20° for all amino acids (except ALA, GLY, and PRO). The mean and median RMSD are calculated according to Krivov *et al.*[37] by averaging over the averages of the amino acid types. Krivov's method SCWRL4 applies the widely used backbone-dependent Dunbrack

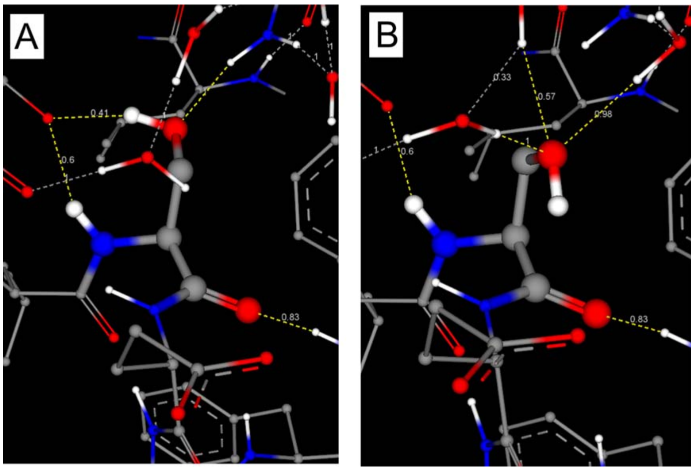


Figure 5

(**A**) Original crystal conformation of SER 935 A of 1KYF, HYDE$_{\text{protein}}$ score: 2.4. (**B**) Conformation found in the remutation, score: −0.3. Water molecules are reoriented by Protoss[44] due to the changed environment. [Color figure can be viewed at wileyonlinelibrary.com]

rotamer library. SRWL4 is trained on 100 protein structures and tested on 379 structures, while our method is not trained on mutational data and tested on 100 high-resolution protein structures (see the section Methods). Using our method, we achieve better RMSD values, while the results for χ-angles below 40° are comparable (Table III). Since Krivov *et al.* did not evaluate the percentage of correct χ-angles below 20°, we cannot compare the performance on this level. However, our results show that 81% of the $\chi1$-angles and 71% of the $\chi1 + \chi2$-angles are below 20°and therefore very close to the original conformation.

Cross-mutation results

The cross-mutation data set contains in total 10,002 protein pairs, of which 1170 have a single mutation, 1833 two-, 865 three-, 4361 four-, and 1773 five-point mutations. In each cross-mutation experiment, all amino acid mutations were analyzed separately. This means that a protein pair with five point mutations results in 10 mutation experiments; that is, every mutation is analyzed individually, thus resulting in five mutations in each direction.

We classified the mutations into three different categories: “single” mutations are those cases where only one amino acid is mutated in the complexes. “multisimple” are those structure pairs, where more than one point mutation occurs. These are most likely harder to predict as the complex might undergo higher conformational changes than with a single point mutation. These multicase mutations are done one by one. Effects of other mutations cannot be seen with our method. The third category are so-called “multihard” cases with more than one mutation and an additional mutation close to the currently investigated mutation site (i.e., within 6.5 Å around the amino acid). In these cases, the complete pocket might change the conformation, thereby making it harder to find the correct conformation. All predictions were performed without crystal water, as these might be replaced by mutations of smaller amino acids into larger ones or are placed at other positions.

Table IV summarizes the overall performance analysis of the cross-mutation experiment. Compared to the remutation experiment, our results here show a decrease in performance with a median RMSD between the original and the predicted side-chain position of 1.15 Å on all cases. Considering the category *single*, only the median RMSD of 0.58 Å is smaller and comparable to the remutation experiment. In the *multisimple* class, the performance decreases to a median RMSD of 0.94 Å and

Table III

Summary of the Remutation Experiments Based on 100 High-Resolution Structures and Comparison to Krivov *et al.*[37] (Results Based on 379 Test Cases)

	With water	Without water	Krivov *et al.*
Mean RMSD	0.56	0.48	0.82
Median RMSD	0.22	0.21	–
% X1 < 40°	85	87	89.3
% X1 < 20°	78	81	–
% X1 + X2 < 40°	77	79	79.7
% X1 + X2 < 20°	68	71	–

Mean RMSD, median RMSD, %$\chi1$ and %$\chi2$ angles between 40° and 20°, with and without water in the complex.

Table IV
Results of the Cross-Mutation Experiment

Category	All	Single	Multisimple	Multihard
Count	7549	956	3051	3542
Median RMSD	1.15	0.58	0.94	1.46
% X1 < 40°	60.97	84.73	70.47	46.39
% X1 < 20°	46.06	72.07	56.93	29.67
% X1 + X2 < 40°	47.56	73.95	57.13	32.19
% X1 + X2 < 20°	35.16	64.85	44.35	19.23

All cases in sum and split into three categories are shown. This evaluation does not contain the cases where only clashing solutions were found. These were in total 1313 of the 8855 cases.

decreases even further for the *multihard* cases with a median RMSD of 1.46 Å. A similar pattern can be observed for the rotational angle performance statistics: for the *single* cases, the $\chi^1 < 40°$ and $< 20°$ are with 85% and 72% almost as good as in the remutation experiment. In case of the *multihard* category, the percentages drop to 46% and 30%, respectively. The same holds true for the comparison between the χ^1 and χ^2-angle difference in the original and the remutated side chains. The performance is comparable to the remutation experiment for the *single* cases (74% and 65%). For the *multihard* cases, percentages of only 32% and 19% were achieved. Note that the data set contains about six times more *multisimple* and *multihard* cases than *single* cases and is therefore biased toward multicases.

For a more detailed analysis of the performance, we analyzed the χ^1 difference smaller 40° for all mutations separately. The results are shown as a heat map in Table V. We find cases which a 100% correctly predicted throughout different types of mutations: from small to larger amino acids like GLY to THR, or VAL to PHE; from hydrophobic to hydrophilic amino acids like LEU to GLU and also from large to small ones like ARG to LEU. This shows that the method is not biased toward one type of mutations. The cases where the method does not find a conformation close to the crystal one also involve different types of mutations (compare Table V). Mutations starting from ALA, GLN, and ARG are best predicted (compare last column of Table V) while conformations of mutations to PHE, TYR, LYS, ASN, and HIS were identified most successfully (compare last row Table V).

Table VI shows the mutation pairs for which only clashing solutions were found by the "rigid" mutation method. Unsurprisingly, mutations starting from smaller amino acids such as GLY and ALA are more likely to produce clashing solutions only than starting with large amino acids such as ARG and LYS. For some mutation pairs such as ALA to TRP or GLY to ASP, all occurrences in the data set result in clashing conformations.

As the backbone RMSD between the two proteins of a mutation pair in the data set can be at most 3.0 Å (SIENA filtering criterion), the reason for clashes are more likely induced by side-chain rearrangements near the mutated residue. Therefore, we evaluated if our method is able to find nonclashing solutions for these cases when all amino acids in the proximity of the

Table V
Percentages of Amino Acid Mutations, Where χ^1 Angle Differences are Below 40°. [Color table can be viewed at wileyonlinelibrary.com]

	Starting Amino acid																	
Target Amino acid	**SER**	**THR**	**CYS**	**VAL**	**LEU**	**ILE**	**MET**	**PHE**	**TYR**	**TRP**	**ASP**	**GLU**	**ASN**	**GLN**	**HIS**	**LYS**	**ARG**	**SUM**
ALA	28.5	94.1	87.9	17.0	100.0	78.0	0.0	44.7	80.0	-	83.3	65.6	55.6	100.0	100.0	100.0	100.0	70.9
GLY	100.0	100.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0	66.7
SER		100.0	11.5	0.0	-	-	-	-	100.0	-	-	0.0	-	-	-	-	0.0	35.3
THR	85.7		58.8	64.9	66.7	75.9	-	-	-	-	0.0	-	-	-	-	-	-	58.7
CYS	39.6	0.0		-	-	-	-	-	100.0	-	-	-	-	-	-	-	-	46.5
VAL	70.6	7.3	-		38.5	63.7	83.3	100.0	-	-	79.2	0.0	-	-	0.0	12.5	-	45.5
LEU	7.7	0.0	100.0	4.1		47.9	99.5	100.0	-	-	50.0	100.0	100.0	100.0	6.8	100.0	100.0	65.4
ILE	100.0	73.1	-	35.6	15.4		64.0	100.0	-	-	0.0	90.1	72.6	50.0	-	-	-	60.1
MET	-	10.2	-	31.6	86.9	0.0		-	-	-	-	-	-	49.7	-	100.0	88.0	52.3
PHE	0.0	-	100.0	0.0	99.0	98.8	0.0		97.9	62.5	-	-	-	-	100.0	-	0.0	55.8
TYR	0.0	51.2	31.3	-	100.0	0.0	-	98.4		-	0.0	-	-	-	-	-	-	40.1
TRP	-	-	-	-	33.3	-	-	81.6	-		-	-	-	-	-	-	0.0	38.3
ASP	-	0.0	-	9.1	100.0	0.0	-	-	-	-		72.7	97.9	100.0	-	-	-	54.2
GLU	0.0	-	50.0	-	100.0	93.1	-	-	-	-	56.3		-	86.3	100.0	-	0.0	60.7
ASN	33.3	-	-	-	-	100.0	-	-	-	-	100.0	-		66.7	50.0	-	-	70.0
GLN	-	50.0	-	-	100.0	33.3	96.5	-	-	-	40.0	76.2	-		100.0	-	100.0	74.5
HIS	-	-	-	27.3	5.3	-	-	100.0	-	-	80.7	40.0	79.0	95.7		-	100.0	66.0
LYS	-	-	-	-	100.0	-	0.0	-	-	-	-	-	-	-	-		50.0	50.0
ARG	-	-	-	-	100.0	-	87.5	-	-	-	-	0.0	-	-	-	100.0		71.9
SUM	42.3	44.2	62.8	21.0	74.6	53.7	53.9	89.3	94.5	62.5	48.9	49.4	81.0	81.0	65.3	82.5	48.9	

Table VI

Percentages of Amino Acid Mutations, Where Only Clashing Solutions are Found. [Color table can be viewed at wileyonlinelibrary.com]

	Starting Amino acid																
Target Amino acid	**SER**	**THR**	**CYS**	**VAL**	**LEU**	**ILE**	**MET**	**PHE**	**TYR**	**TRP**	**ASP**	**GLU**	**ASN**	**GLN**	**HIS**	**LYS**	**ARG**	**AVG**
ALA	2.7	5.6	12	71.4	32	0	77	53.5	28.6	100	0	46	63	94	0	0	47	37.2
GLY	77	97		100							100	100	100	100			48	90.2
SER		2.3	17	33.3				100	0	100		0					0	31.6
THR	0		0	5.13	0	15			100		0							17.2
CYS	0	0							92.6									30.9
VAL	0	0			0	0	0	70			0	0			75	33		17.8
LEU	0	0	0	14.7		1	0	62.2	100	100	0	0	0	0	20	0	0	18.6
ILE	0	0		0	0		0	1.19		100	0	0	0	0				9.2
MET		16		13.6	0.4	4								0		0	0	4.8
PHE	0		0	60	0	0	0		7.69	15.8					0		0	8.3
TYR	0	0	0		0	0		15.8			0							2.3
TRP					0			5									0	1.7
ASP		0		0	0	7						0	2	20				4.2
GLU	0		0		0	1					0			0	33.3		0	4.3
ASN	0					11					0			0	33.3			8.9
GLN		0			0	0	0				0	0			83.3		0	10.4
HIS				0	0			75			0	0	0	0			0	9.4
LYS					0		0										0	0.0
ARG					0		0					0				0		0.0
AVG	7.3	11.0	4.2	29.8	2.3	3.6	9.6	47.8	54.8	83.2	9.1	14.6	27.4	23.8	35.0	6.7	8.6	

mutation are flexible upon optimization. In summary, 844 of these 1313 clashing cases could be resolved (64%). However, introducing more flexibility also allows the optimizer to explore alternative local minima, making it harder to identify the global minimum (crystal structure conformation). Of the 844 conformations, only 208 (25%) have an RMSD smaller than 1 Å compared to the aligned crystal structure. On the other hand, the χ-angle performance is more promising, with 59% of the χ^1 angle differences below 40° (58% below 20°) and 32% of $\chi^1 + \chi^2$ angle differences below 40° (24% below 20°). The RMSD performance is not as good as the χ^1 angle performance since the alignment is a global one and even small backbone changes might result in a not exact side-chain alignment.

Finally, we show two detailed examples of successfully resolved clashes upon cross-mutation (Fig. 6). In Figure

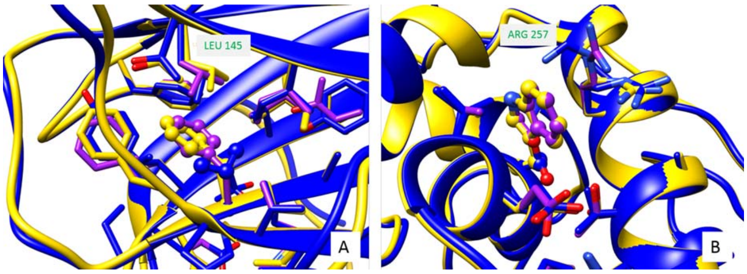


Figure 6

Examples of remedy of clashes by flexibility of amino acid side chains. In yellow, is shown the mutated protein; in blue, the nonmutated structure; in purple, the conformations of the fully flexible pocket side chains are depicted. (**A**) Yellow: 4KVP with PHE 157. Blue: 3D06 with VAL 157. Purple: mutated residue PHE 157 and results of side-chain optimization. (**B**) Yellow: 4FCS with TRP 164. Blue: 3FMU with SER 164. Purple: mutated residue TRP 164 and results of side-chain optimization. [Color figure can be viewed at wileyonlinelibrary.com]

Table VII
Results of the Stability Prediction Experiment by Evaluating the Correctness of Stabilizing Versus Destabilizing Effects on the Data Set of Khan *et al.*[24]

Subset	# mutations	TP	FP	TN	FN	Acc	TNR	TPR	MCC
All	1069	125	325	562	57	0.64	0.63	0.69	0.24
Buried	336	24	73	222	17	0.73	0.75	0.59	0.24
Medium resolution	701	95	196	370	40	0.66	0.65	0.70	0.29
High resolution	220	35	65	103	17	0.63	0.61	0.67	0.24
High-resolution buried	85	13	23	44	5	0.67	0.66	0.72	0.31

# mutations, number of mutations; TP, number of true positives (stabilizing mutations); TN, number of true negatives (destabilizing mutations); FP, false positives; FN, false negatives; Acc, accuracy; TNR, true negative rate/specificity; TPR, true positive rate/sensitivity.

Table VIII
Results of the Stability Prediction Experiment With Clashes Removed Due to a Fully Flexible Site

Subset	# mutations	TP	FP	TN	FN	Acc	TNR	TPR	MCC
All	1069	135	354	533	47	0.62	0.60	0.74	0.26
Buried	336	26	88	207	15	0.69	0.70	0.63	0.23
Medium resolution	701	103	217	349	32	0.64	0.62	0–76	0.30
High resolution	220	40	71	97	12	0.62	0.58	0.77	0.29
High-resolution buried	85	13	26	41	5	0.63	0.61	0.72	0.27

# mutations, number of mutations; TP, number of true positives (stabilizing mutations); TN, number of true negatives (destabilizing mutations); FP, false positives; FN, false negatives; Acc, accuracy; TNR, true negative rate/specificity; TPR, true positive rate/sensitivity.
Evaluation of the correctness of stabilizing versus destabilizing effects on the data set of Khan *et al.*[24]

6(A), an oncogenic mutant (VAL157PHE) of the tumor suppressor protein p53 is superimposed with the original nonmutated structure. As a third structure, the conformations identified by keeping the side chains flexible are depicted. In the nonmutated structure, LEU 145 is rotated into the pocket and thus would clash with PHE 157 if it would be kept rigid during the mutation process. In the optimized pocket (purple), it is moved out of the way while the other side chains are kept in place. The RMSD of the mutated conformation of PHE compared to the crystal structure is 0.68 Å. In Figure 6(B), a mutation (SER164TRP) of a versatile peroxidase is shown (yellow: PDB 4FCS, TRP 164; blue: PDB 3FMU, SER 164). The side-chain causing the clash here is ARG 257. It seems to be very flexible at this position, as in the 3FMU crystal structure, two alternating conformations are given. In the mutation experiment replacing the SER with TRP and allowing all amino acid side chains to rotate, this ARG adapts a different conformation (purple). The conformation found (purple) differs from the crystal structure, but allows the TRP to achieve the same conformation as in the crystal structure (RMSD: 0.13 Å).

Energy-prediction results

In the energy-prediction experiment, we assess the methods capabilities to predict the effect of a mutation on the overall energy of a protein. Three types are possible: stabilizing, destabilizing, or neutral mutations. In the first experiment, we omitted the neutral cases and assessed if our method correctly identifies stabilizing and

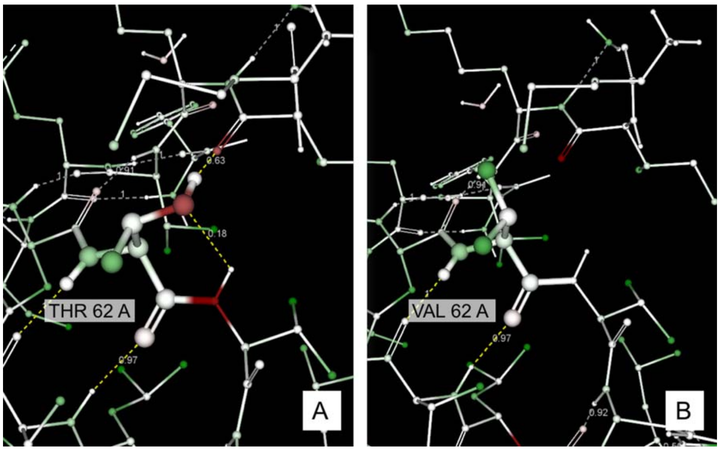


Figure 7
A favorable mutation of THR 62 A into VAL 62 A of a DNA-binding protein (PDB code 1VQB). (**A**) Energetically unfavorably contributing original amino acid side-chain scored with HYDE$_{\text{protein}}$; red = unfavorable energy contributions; green = favorable energy contribution. (**B**) Mutated side chain with mainly favorable energy contribution. [Color figure can be viewed at wileyonlinelibrary.com]

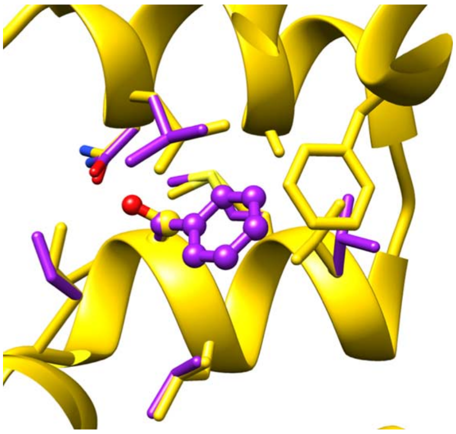


Figure 8

Cross-mutation with fully flexible pocket amino acids. Mutation of SER A 117 to PHE of a T4 lysozyme (PDB code 2LZM). Original crystal structure in yellow; cross-mutated structure in purple. To accommodate the larger PHE amino acid, LEU 121 and LEU 133 are moved out of the way. [Color figure can be viewed at wileyonlinelibrary.com]

destabilizing mutations. A mutation is classified as destabilizing if the differences in the HYDE$_{\text{protein}}$ score of the original versus the mutated amino acid are positive and stabilizing otherwise. As we are relying on high-quality crystal structures in the HYDE scoring method due to the detailed geometric modeling of hydrogen bonds, we classified the data set provided by Khan *et al.*[24] and only evaluated buried mutations (see the section "Methods"). Table VII shows the results of this experiment. Matthew's correlation coefficient is with 0.31 best for the high-resolution buried cases, followed by 0.29 for the medium-resolution cases, indicating a marginal better performance for buried structures. The correlation is within the range of other methods,[24] although the correlation coefficients are not directly comparable, due to the usage of different data sets for the evaluation of the other methods.

Figure 7 shows an example for a stabilizing mutation. According to ProTherm data, mutating THR A 62 of a DNA binding protein (PDB code 1VQB) to VAL stabilizes the protein with a $\Delta\Delta G$ of 1.3 kJ/mol. In Figure 7(A), the original amino acid is highlighted in the HYDE coloring scheme and shows mainly energetically unfavorable contributions (red). The amino acid side-chain is involved in unfavorable interactions with ILE-A-2 and VAL-A-63 in the crystal structure, which leads to a penalty in the HYDE scoring function. The total score of the amino acids is with 0.86 unfavorable. Figure 7(B) shows the mutation to the hydrophobic VAL. This resolves the energetically unfavorable situation: not only the amino acid itself is scored much better (HYDE-score of −5.7) but also the surrounding amino acids achieve a favorable score of −13.5. This stabilizing mutation was found by Sandberg *et al.*,[51] who also evaluated the activity of the protein mutants. They found that this mutation is stabilizing but not improving the activity of DNA binding. This illustrates a common effect: stabilization does not necessarily mean that the mutant is also more active.

Table IX

Percentage of Correct Predicted Mutation Effects by HYDE$_{\text{protein}}$ for Stabilizing, Destabilizing, and Neutral Cases on the Data Set of Khan *et al.*[24]

	% Correct			
	Stabilizing	Destabilizing	Neutral	Total
All	55.49	54.23	25.61	43.52
Buried	53.66	71.86	12.03	53.30

In the experiment as described above, a mutation resulting in only clashing conformations of the mutated amino acid side chain was counted as destabilizing, despite its experimentally validated effect. In a second experiment, we re-evaluated these cases allowing the surrounding amino acids to be flexible, thus the pocket side chains can be rotated during the optimization (Table VIII). The HYDE$_{\text{protein}}$ score is used to determine the best conformation. Of the original 94 cases with clashing solutions (83 destabilizing and 13 stabilizing), 50 could be resolved using this setup. Only three structures remained for which only clashing solutions were found and the mutation was classified as stabilizing. Unfortunately, for those mutations, no crystal structures are available in the PDB. Therefore, we could not investigate how the protein structures accommodate the mutated amino acids. It might be that in these cases, a backbone movement is needed, as the amino acids are in rather tight pockets.

Figure 8 shows one example where the site of the amino acid mutation is in a rigid pocket and only clashing solutions are found even though the experimental $\Delta\Delta G$ is favorable. Anderson *et al.*[52] reported a stabilizing mutation of SER A 117 to PHE A 117 of a T4 lysozyme (PDB code 2LZM), which leads to a higher thermostability of the protein (4.8°C higher melting point). Keeping all amino acids in the pocket flexible, a nonclashing mutation is achieved due to rearranging LEU A 121 and LEU A 113; a rearrangement also described by Anderson *et al.* Achieving a HYDE-score of −3.5, the hydrophobic PHE is scored much better in its hydrophobic environment than the original SER. Owing to an unsatisfied oxygen atom pointing into the hydrophobic pocket, SER got an unfavorable HYDE-score of 4.4.

Finally, we included neutral mutations as a potential outcome and evaluated the capability of the method to distinguish these three mutation effects: stabilizing, destabilizing, or neutral. For neutral cases, we used an HYDE$_{\text{protein}}$ score of ±1 as a cutoff. The results are shown in Table IX: for each class, the percentage of correctly predicted cases is given, and the total percentage of correctly predicted cases. The evaluation shows that in this three-class prediction, our method better predicts stabilizing and destabilizing effects: 54% and 72% cases are predicted correctly for the buried subset, respectively. However, the neutral cases are only in 12–26% predicted correctly. In general, buried amino acids are more likely to be correctly predicted (compare Table IX). This evaluation shows that a prediction method such as HYDE$_{\text{protein}}$ should be used to predict strong effects like clearly stabilizing or destabilizing.

CONCLUSION

In this work, we focused on two major aspects: (1) the evaluation of HYDE$_{\text{protein}}$ to predict protein mutations and (2) the introduction of a thorough evaluation strategy including a benchmarking data set with protein mutation pairs.

We assessed the performance of HYDE$_{\text{protein}}$ to predict protein structure changes in the context of mutations. The most basic task is to predict the side-chain conformation of an amino acid within its native structural environment. Assessing this task, we found that HYDE$_{\text{protein}}$ is performing comparably to the other state of the art protein conformation prediction tools. In contrast to these, HYDE$_{\text{protein}}$ is a physically motivated scoring function without any supervised training to mutation data.

In the second part of this publication, we introduced a novel cross-mutation performance experiment and a new data set. This experiment is in our opinion more difficult and represents a more realistic application scenario than the simple remutation experiment. Not surprisingly, a performance drop can be found in this setup. We would like to encourage other researchers to use this experiment in combination with the data set for their performance evaluation (data set can be found in the Supporting Information).

In the third experiment, we evaluated the ability of HYDE$_{\text{protein}}$ to predict the energetic effect of a mutation. Herein, HYDE$_{\text{protein}}$ shows promising results as soon as the mutational effect is clearly stabilizing or destabilizing. Future development will go into the correct identification of energetically neutral effects that are more difficult to predict. The results are especially encouraging as HYDE has not been trained on mutational data and thus exploits the physical basis to predict protein stability and provides a deeper insight into protein stability. Therefore, we are convinced that with HYDE$_{\text{protein}}$, we are on a promising way toward predicting mutation effects.

HYDE$_{\text{protein}}$ could be further enhanced by considering the alternative conformations of amino acids given in the PDB files. In our study, consistently, the first conformation listed in the PDB file was chosen for analysis. However, the second conformation might lead to better results. Furthermore, mutations into CYS are currently not analyzed concerning their ability to form disulfide bridges. Thus, mutations to CYS might give wrong results. Another potential enhancement could be the usage of conformation libraries rather than applying an unguided enumeration strategy as these libraries have been shown to improve current methods.[34–37]

Two effects that are neglected by most mutation prediction methods are backbone flexibility and the correct handling of water molecules. Both aspects are very complex and not easy to solve. However, they could lead to a great enhancement of the method's accuracy. Especially, backbone flexibility will certainly be necessary to correctly predict mutations that lead to a greater conformational change. Currently, water molecules are represented implicitly in HYDE, if no χ-ray observed ones are present. However, the correct placement and integration of water molecules displays a further aspect of improvement.

Compared to methods based on the assumption that structural observations are energetically favorable, HYDE$_{\text{protein}}$ can even be applied if no previous information is available. As the contribution of molecular interactions is modeled on a pure physical basis, a prediction of underrepresented or complete missing mutations in available data will be possible. HYDE$_{\text{protein}}$ is showing promising results on mutation conformation prediction while it could profit from relevant finding in the field like conformational libraries.

AUTHOR CONTRIBUTIONS

KS has written the manuscript, developed and conducted all the experiments, and contributed to the concept of HYDE$_{\text{protein}}$. EN has contributed to the manuscript, assembled the data set for the energy-prediction experiment, and contributed to the design of the experiments. AM has contributed to the manuscript and implemented the HYDE optimizer. SB has contributed to the manuscript and helped implementing the HYDE mutation strategy. NS has implemented the prototype of the HYDE$_{\text{protein}}$ scoring method and initiated together with GL, RK, and MR, the concept of HYDE$_{\text{protein}}$. MR has supervised the project. All authors have read the final manuscript and given their approval to it.

ACKNOWLEDGMENT

The authors want to thank Christian Lemmen and the BioSolveIT for helpful discussions during the project.

REFERENCES

1. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K. Engineering the third wave of biocatalysis. Nature 2012; 485:185–194.

2. Bornscheuer UT, Pohl M. Improved biocatalysts by directed evolution and rational protein design. Current Opinion in Chemical Biology 2001;5:137–143.
3. Eijsink VGH, Gåseidnes S, Borchert T V., Van Den Burg B. Directed evolution of enzyme stability. Biomolecular Engineering 2005;22:21–30.
4. Chen R. Enzyme engineering: Rational redesign versus directed evolution. Trends in Biotechnology 2001;19:13–14.
5. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, Meiler J. Practically useful: What the R osetta protein modeling suite can do for you. Biochemistry 2010;49:2987–2998.
6. Magliery TJ. Protein stability: Computation, sequence statistics, and new experimental methods. Current Opinion in Structural Biology 2015;33:161–168.
7. Shirke AN, Basore D, Butterfoss GL, Bonneau R, Bystroff C, Gross RA. Toward rational thermostabilization of Aspergillus oryzae cutinase: Insights into catalytic and structural stability. Proteins: Structure, Function and Bioinformatics 2016;84:60–72.
8. Jha RK, Chakraborti S, Kern TL, Fox DT, Strauss CEM. Rosetta comparative modeling for library design: Engineering alternative inducer specificity in a transcription factor. Proteins: Structure, Function and Bioinformatics 2015;83:1327–1340.
9. Ng PC, Henikoff S. Predicting the Effects of Amino Acid Substitutions on Protein Function. Annual Review of Genomics and Human Genetics 2006;7:61–80.
10. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins: Structure, Function, and Bioinformatics 2005;62:1125–1132.
11. Masso M, Vaisman II. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics 2008;24:2002–2009.
12. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: Prediction of protein stability upon point mutations. Nucleic Acids Research 2006;34:W239–W242.
13. Parthiban V, Gromiha MM, Hoppe C, Schomburg D. Structural analysis and prediction of protein mutant stability using distance and torsion potentials: Role of secondary structure and solvent accessibility. Proteins: Structure, Function and Genetics 2007;66:41–52.
14. Tian J, Wu N, Chu X, Fan Y. Predicting changes in protein thermostability brought about by single- or multi-site mutations. BMC Bioinformatics 2010;11:370.
15. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science 2009;11:2714–2726.
16. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. Journal of Molecular Biology 2002;320:369–387.
17. Gilis D, Rooman M. PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins. Protein Engineering Design and Selection 2000;13:849–856.
18. Capriotti E, Fariselli P, Casadio R. A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 2004;20:i63–i68.
19. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research 2005;33:W306–W310.
20. Deutsch C, Krishnamoorthy B. Four-body scoring function for mutagenesis. Bioinformatics 2007;23:3009–3015.
21. Dosztányi Z, Magyar C, Tusnády GE, Simon I. SCide: Identification of stabilization centers in proteins. Bioinformatics 2003;19:899–900.
22. Magyar C, Gromiha MM, Pujadas G, Tusnády GE, Simon I. SRide: A server for identifying stabilizing residues in proteins. Nucleic Acids Research 2005;33:W303–W305.
23. Dosztányi Z, Fiser A, Simon I. Stabilization centers in proteins:Identification, characterization and predictions. Journal of Molecular Biology 1997;272:597–612.
24. Khan S, Vihinen M. Performance of protein stability predictors. Human Mutation 2010;31:675–684.
25. Bava KA. ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Research 2004;32:120D–121D.
26. Sarai A. ProTherm. http://www.abren.net/protherm/protherm.php [accessed 2017 Jan 4].
27. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. Protein Engineering, Design and Selection 2009;22:553–560.
28. Benedix A, Becker CM, de Groot BL, Caflisch A, Böckmann RA. Predicting free energy changes using structural ensembles. Nature Methods 2009;6:3–4.
29. Pokala N, Handel TM. Energy functions for protein design: Adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. Journal of Molecular Biology 2005;347:203–227.
30. Rohl C, Strauss C, Misura K, Baker D. Protein structure prediction using Rosetta. Methods in enzymology 2003;383:66.
31. Potapov V, Cohen M, Inbar Y, Schreiber G. Protein structure modelling and evaluation based on a 4-distance description of side-chain interactions. BMC Bioinformatics 2010;11:374.
32. Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. MAESTRO - multi agent stability prediction upon point mutations. BMC Bioinformatics 2015;16:116.
33. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. Journal of Molecular Biology 2001;311: 421–430.
34. Dunbrack RL. Rotamer libraries in the 21st century. Current Opinion in Structural Biology 2002;12:431–440.
35. Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. Protein Science 2003; 12:2001–2014.
36. Wang Q, Canutescu AA, Dunbrack RL. SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. Nature Protocols 2008;3:1832–1847.
37. Krivov GG, Shapovalov M V., Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Structure, Function and Bioinformatics 2009;77:778–795.
38. Subramaniam S, Senes A. An energy-based conformer library for side chain optimization: Improved prediction and adjustable sampling. Proteins: Structure, Function and Bioinformatics 2012;80: 2218–2234.
39. Liang S, Grishin N V. Side-chain modeling with an optimized scoring function. Protein Science 2009;11:322–331.
40. Petrella RJ, Lazaridis T, Karplus M. Protein sidechain conformer prediction: A test of the energy function. Folding and Design 1998; 3:353–377.
41. Schneider N, Lange G, Hindle S, Klein R, Rarey M. A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: Methods behind the HYDE scoring function. Journal of Computer-Aided Molecular Design 2013;27: 15–29.
42. Schärfer C, Schulz-Gasch T, Ehrlich HC, Guba W, Rarey M, Stahl M. Torsion angle preferences in druglike chemical space: A comprehensive guide. Journal of Medicinal Chemistry 2013;56: 2016–2028.
43. Guba W, Meyder A, Rarey M, Hert J. Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. Journal of Chemical Information and Modeling 2016;56:1–5.
44. Bietz S, Urbaczek S, Schulz B, Rarey M. Protoss: A holistic approach to predict tautomers and protonation states in protein-ligand complexes. Journal of Cheminformatics 2014;6:12.

45. Hansch, C, Leo, A, Hoekman, D. Exploring QSAR: Hydrophobic, Electronic, and Steric Constants. American Chemical Society, Washington, DC. 1995.
46. Schneider N, Volkamer A, Nittinger E, Rarey M. Supporting biocatalysis research with structural bioinformatics. In: Applied Biocatalysis: From Fundamental Science to Industrial Applications. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA; 2016. p. 71–100.
47. Nittinger E, Schneider N, Lange G, Rarey M. Evidence of water molecules - A statistical evaluation of water molecules based on electron density. Journal of Chemical Information and Modeling 2015;55:771–783.
48. Pottel J, Moitessier N. Single-Point Mutation with a Rotamer Library Toolkit: Toward Protein Engineering. Journal of Chemical Information and Modeling 2015;55:2657–2671.
49. Bietz S, Rarey M. SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles. Journal of Chemical Information and Modeling 2016;56:248–259.
50. Bietz S, Rarey M. ASCONA: Rapid Detection and Alignment of Protein Binding Site Conformations. Journal of Chemical Information and Modeling 2015;55:1747–1756.
51. Sandberg WS, Schlunk PM, Zabin HB, Terwilliger TC. Relationship between in Vivo Activity and in Vitro Measures of Function and Stability of a Protein. Biochemistry 1995;34:11970–11978.
52. Anderson DE, Hurley JH, Nicholson H, Baase WA, Matthews BW. Hydrophobic core repacking and aromatic-aromatic interaction in the thermostable mutant of T4 lysozyme ser 117–phe. Protein Sci. 1993;2:1285–1290.

D.2 Manuscripts in Preparation

Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples.

[P1] **Nittinger, E.**; Flachsenberg, F.; Bietz, S.; Lange, G.; Rarey, M. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. J. Chem. Inf. Model. 2018, *Accepted*.

`http://pubs.acs.org/articlesonrequest/AOR-AWK5Q7nVyVFIr7ztgyds`

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

pubs.acs.org/jcim

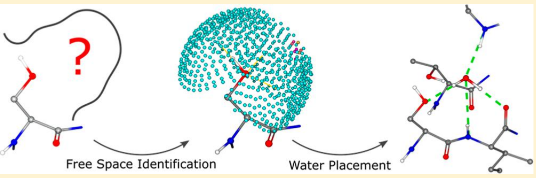
Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples

Eva Nittinger,[†] Florian Flachsenberg,[†] Stefan Bietz,[†] Gudrun Lange,[‡] Robert Klein,[‡] and Matthias Rarey*,[†]

[†]Universität Hamburg, ZBH − Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg, Germany

[‡]Bayer CropScience AG, Industriepark Hoechst G836, 65926 Frankfurt am Main, Germany

Supporting Information



ABSTRACT: Water molecules are of great importance for the correct representation of ligand binding interactions. Throughout the last years, water molecules and their integration into drug design strategies have received increasing attention. Nowadays a variety of tools are available to place and score water molecules. However, the most frequently applied software solutions require substantial computational resources. In addition, none of the existing methods has been rigorously evaluated on the basis of a large number of diverse protein complexes. Therefore, we present a novel method for placing water molecules, called WarPP, based on interaction geometries previously derived from protein crystal structures. Using a large, previously compiled, high-quality validation set of almost 1500 protein−ligand complexes containing almost 20 000 crystallographically observed water molecules in their active sites, we validated our placement strategy. We correctly placed 80% of the water molecules within 1.0 Å of a crystallographically observed one.

INTRODUCTION

A good understanding of water molecules and their interactions with proteins and small molecules is essential for the prediction of protein−ligand binding geometries and affinities. Not surprisingly, the interest in individual water molecules and their contribution to molecular interactions, and by this to the binding affinity, has increased dramatically in the past years. Not only do water molecules mediate interactions between protein and ligands, but also, their displacement can be a major contributor to protein−ligand binding affinity.[1,2]

This increased interest is also reflected by the number of tools and methods available nowadays for the prediction, placement, and scoring of water molecules, ranging from rather simple geometric scoring criteria to extensive molecular dynamics (MD) simulations. Available methods can be separated into four different classes: (1) empirical and knowledge-based methods (Consolv,[3] WatCH,[4] WaterScore,[5] PyWATER,[6] Proasis WaterRank,[7] the relevance metric,[8] AQUARIUS2,[9,10] WATGEN,[11] AcquaAlta,[12] WaterDock,[13] Tetrahedron-water-cluster model,[14] Fold-X,[15] HINT (Hydropathic Interactions) toolkit,[16] Dowser++[17]), (2) statistical and molecular mechanics methods (GRID,[18−20] 3D-RISM,[21,22] MCSS,[23−25] WaterFLAP,[26] wPMF,[27] SZMAP[28]), (3) MD simulation methods (WaterMap,[29,30] GIST,[31] STOW,[32] WATCLUST,[33] SPAM,[34] WATsite,[35,36] GCT,[37−39] BiKi Hydra[40]), and (4) Monte Carlo simulation methods (RETI,[41,42] the double decoupling method,[43,44] double decoupling with RETI,[45] MCRS,[46] JAWS,[47] GCMC[48−50]). The first category can be further classified according to the aim of the method. Some of those methods identify conserved crystallographically determined water molecules, others try to assign a relevance score to them, while others place and/or score water positions. The number of protein structures used for evaluation of the methods declines throughout the four classes. Empirical and knowledge-based methods have been evaluated on seven to 193 structures, statistical and molecular mechanics methods on zero to 100 structures, MD simulation methods on fewer than 10 structures, and Monte Carlo simulation methods on fewer than 15 structures. Table 1 provides a comprehensive list of all of these methods, including short descriptions and information about their evaluation. We refer to a recent review[51] and perspective[52] on water for more detailed information about the various methods.

A consistent, reliable, and fast water placement procedure is important for different application scenarios. Crystal structures with low resolution (>2.7 Å) do not allow modeling of water molecules.[69] Usually, protein−ligand docking poses are generated without water molecules. However, water molecules are important for the correct estimation of their binding affinity. Most of the frequently used software solutions for water placement are time-consuming, preventing their dynamic application to a large number of protein−ligand complexes.

Especially for the development of water placement and prediction methods, the data used for training as well as evaluation display an important and at the same time difficult aspect. The individual energy contribution of a water molecule cannot be measured experimentally. The difference in energy

Received: May 4, 2018

DOI: 10.1021/acs.jcim.8b00271
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Summary of Different Water Scoring and Placement Methods (H-Bond = Hydrogen Bond; IFST = Inhomogeneous Fluid Solvation Theory)

method	description	evaluation
Empirical/Knowledge-Based Methods: Identification of Conserved X-ray Water Molecules		
Consolv[3]	evaluation of four environmental factors: *B* factor, number of H-bonds to the protein, density of neighboring protein atoms, hydrophilicity	training: 13 free vs ligand bound structures; test: seven structures (75%); 1.2 Å between waters in the ligand-free and bound structures as the conservation criterion
WatCH[4]	hierarchical clustering to identify conserved waters in related structures	10 thrombin, three trypsin, four BPTI, two trypsin/BPTI structures
WaterScore[5]	score consisting of a combination of *B* factor, solvent contact surface area, total H-bond energy, and number of protein contacts <3.5 Å	training: 25 protein pairs; 0.5 Å between waters in the ligand-free and bound structures as the conservation criterion
PyWATER[6]	use of structural information; water quality assessment by calculation of two normalized *B* factors: mobility[4] and Carugo's normalized atomic displacement factor[53]	data from WatCH;[4] 12 MHC class I protein structures and a bromodomain-containing protein 4; "PyWATER identified all water molecules"
Empirical/Knowledge-Based Methods: Relevance of X-ray Water Molecules		
Proasis WaterRank[7]	geometric score for the displaceability of waters; distances and angles for a maximum of two donors and two acceptors around the water	N/A
relevance metric[8]	combination of WaterRank[7] and HINT[54]	training: 25 protein pairs (86%); test: 16 protein pairs (87%); 1.2 Å between waters in the ligand-free and bound structures as the conservation criterion
Empirical/Knowledge-Based Methods: Placement of (Structurally Relevant) Water Molecules		
AQUARIUS2[9,10]	prediction of water positions on the basis of known water positions around the 20 amino acids; likelihood weights calculated for grid points	distributions of water around 16 structures; test: five structures; 66% placed waters within 1.0 Å of an X-ray-observed water
WATGEN[11]	water site prediction in protein−protein and protein−peptide interfaces on the basis of experimentally observed geometries of hydration around amino acids	126 interfaces, 72% within 1.5 Å
AcquaAlta[12]	prediction of protein−ligand bridging waters on the basis of data derived from the CSD;[55] ranking of water positions on the basis of ab initio calculations	20 crystal structures with 76% within 1.4 Å; test: 14 complexes with 66%
WaterDock[13]	use of AutoDock Vina[56] for the prediction of ordered waters; probabilistic water molecule classifier to predict conservation of water sites and identify type of displacement (i.e., polar/apolar)	training: 91 consensus waters, 88% within 2.0 Å, 79% within 1.5 Å, 59% within 1.0 Å; test: AcquaAlta[12] set, 87% within 1.4 Å; test conservation prediction: 1004 waters, 75% accurate; test displacement prediction: 459 waters, 80% accurate
tetrahedron-water-cluster model[14]	derivation of feature triangles based on residue triplets; three polar atoms form the bottom triangle of a tetrahedron, and the water molecule is placed at its top	training: 2003 complexes with <2.0 Å resolution to derive triangles; test: 193 complexes, water molecules with interactions to three amino acid residues and the ligand, 73% below 1.5 Å
Empirical/Knowledge-Based Methods: Placement of Water Molecules and Energy Estimation		
Fold-X[15]	prediction of waters in biomolecular structures by the AQUARIUS approach;[18] incorporation of positions in the Fold-X force field	50 crystal structures, 76% with an average RMSD of 0.8 Å; integration of waters into energy calculations for protein mutation effect (overall no significant effect)
HINT (Hydropathic Interactions) toolkit[16]	identification of free grid points and subsequent scoring with HINT;[54] placement of water molecules at most favorable positions; HINT force field estimates Gibbs free energies on the basis of vdW interactions and partial atomic partition coefficients	overall accuracy of 1.28 ± 0.55 Å based on 23 HIV structures; comparison with GRID: 43/101 placed waters within the corresponding GRID contours
Dowser++[17]	Dowser[57] plus AutoDock Vina[56] and WaterDock;[13] search and filling of internal cavities detected with a water probe of 1 Å	14 structures, 85% predicted within 2.0 Å
Statistical/Molecular Mechanics Methods: Prediction of Water Positions and/or Energy Estimation		
GRID[18−20]	aim: generation of contour surfaces to identify energetically favorable positions for the respective probes; location of favorable water positions using a chemical probe mimicking a water (among other probes); the entropic contribution is estimated indirectly by evaluating the hydrophobic effect with a "dry" probe	different studies applying GRID to place waters,[58−60] i.e., 88 X-ray waters of cytochrome *c* oxidase are located in energetically favorable positions[61]
3D reference interaction site model (3D-RISM)[21,22]	aim: generation of 3D solvent site profiles; integral-equation theory of liquids; the approach allows equilibrium solvent distributions to be obtained rapidly without sampling, and thus, no insight into the dynamics is possible; the distributions reveal favorable hydration sites with localized entropies, enthalpies, and solvation free energies; full atomistic sampling of solvent	placement:[62] average error for placement of 11 waters = 0.65 ± 0.24 Å
multiple-copy simultaneous search (MCSS)[23−25]	random distribution of probes (i.e., a water probe) followed by energy minimization	N/A
WaterFLAP[26]	GRID-based water prediction with different probes for entropic contribution (CRY, ENTR)	seven targets, ~90% within 1.5 Å, ~60% within 1.0 Å
water potential of mean force (wPMF)[27]	aim: predict hydration sites in proteins; radial distribution functions of water in the proximity of protein atom types combined with equivalent potentials of mean force to predict hydration sites and assign wPMF scores to waters; trained on 3946 protein structures: extraction of water structure pattern and hydrophilicity; grid-based clustering scheme with wPMF to predict water sites	100 crystal structures; 80% of predicted clusters occupied by an X-ray water within 1.4 Å

Table 1. continued

method	description	evaluation
Statistical/Molecular Mechanics Methods: Prediction of Water Positions and/or Energy Estimation		
solvent Zap mapping (SZMAP)[28]	grid-based semicontinuum approach; one explicit water molecule in combination with a Poisson−Boltzmann continuum model	finding X-ray waters: testing on two structures, more than 50% of the X-ray waters in the active site were within a SZMAP n_ddG map at a contour level of −0.5 kcal/mol; conservation classification of 54 waters with 94% accuracy; energy prediction comparison with RETI[45]
Molecular Dynamics Simulation Methods: Prediction of Water Positions and/or Energy Estimation		
WaterMap[29,30]	nanoscale MD simulation followed by IFST[63,64] calculation to determine thermodynamics; waters from MD are clustered to derive water positions	one structure with 11 waters, nine within 1.5 Å; further applications shown in ref 65
grid inhomogeneous solvation theory (GIST)[31]	aim: estimate the solvation free energy, discretization of IFST onto a 3D grid	overlay of GIST contour plots with hydration sites[66]
Solvation Thermodynamics of Ordered Water (STOW)[32]	MD simulation in combination with IFST to provide mean energetic interaction of specific water positions; calculation of entropic penalty by rotational and translational restrictions	N/A
WATCLUST[33]	aim: prediction of water sites for further integration into docking MD simulations to derive water sites and determine structural and thermodynamic properties	N/A
SPAM[34]	"MAPS" spelled backward; nanoscale MD simulation with explicit solvent; calculation of free energy using a site partition function based on the local distribution relative to the perturbation in bulk water; neglect of water−water contacts	energetic evaluation of one HIV water; no placement evaluation
WATsite[35,36]	clustering of MD trajectories; entropy estimations from probability density of translation and rotation of waters; no water−water contacts are considered	N/A
grid cell theory (GCT)[37−39]	aim: calculation of solvent thermodynamics; calculation of thermodynamic properties by comparison of water−solute and water−water interaction energies; computation of solvent thermodynamics by the cell theory method upon molecular dynamics simulation	computed energies of small organic molecules correlate well with TI calculations; comparison of three protein families with MC/FEP calculations: qualitative but not quantitative agreement
BiKi Hydra[40]	aim: analysis of hydration patterns; MD-based analysis of the persistence of water molecules; steered MD simulation followed by spatial density analysis to measure local water stability; differentiation of persistence due to favorable interactions or steric hindrance	water placement validation using the A_{2A} receptor; persistence evaluation of the A_{2A} receptor and four ligands
Monte Carlo Simulation Methods: Prediction of Water Positions and/or Energy Estimation		
replica-exchange thermodynamic integration (RETI)[41,42]	aim: calculation of relative hydration free energies; combination of finite-difference thermodynamic integration (FDTI) and Hamiltonian replica-exchange method	N/A
double decoupling method[43,44]	aim: standard free energy calculation; based on statistical thermodynamics; water decoupled from the bulk using molecular dynamics thermodynamic integration and water decoupled from the receptor using thermodynamic integration (double decoupling)	comparison of calculated energies for decoupling of water from the bulk to gas with results from ref 67
double decoupling with RETI[45]	aim: determination of water energies to predict displaceability; combination of double decoupling[44] with RETI;[41,42] Monte Carlo-based method for calculating binding energies; all waters are treated explicitly; two Monte Carlo simulations are run, i.e., water decoupled from the bulk and water decoupled from the receptor (double decoupling)	correlations between calculated water energies and experimentally measured energy changes upon ligand modification for six protein families (54 water energy calculations)
Monte Carlo reference state (MCRS)[46]	simulation of the expected atom−atom contact density probability by sampling of structural space with random probes using a Monte Carlo-based method	test: 12 protein structures, 50% within 1.0 Å (average RMSD = 1.29 Å); energy prediction tested using fold recognition experiments
Just Add Water Molecules (JAWS)[47]	statistical thermodynamics methodology based on λ-dynamics; conformational sampling of the protein chain while waters appear and disappear (double decoupling) on a grid; possibility to calculate hydration site occupancies that are incorporated into the interaction energy calculation.	five proteins for testing, i.e., scytalone dehydratase with four X-ray waters, and four predicted within 2.0 Å; few comparisons of predicted energies with results from double decoupling[45]
grand canonical Monte Carlo (GCMC)[48−50]	the number of particles (i.e., water molecules) fluctuates depending on a defined chemical potential; the relation between the free energy of water and the chemical potential allows conclusions to be drawn about the water affinity at specific locations	further development and testing:[68] mouse major urinary protein I as a negative control, five protein structures from WaterDock,[13] 100% within 2.0 Å

measured upon displacement of a water molecule is always in combination with the extension of the ligand. Additionally, the experimental evidence for a specific location of a water molecule should be included. An ideal data set would contain not only the oxygen atom position of the water molecule but also the orientations of the hydrogen atoms. Those "ideal" data could be retrieved from neutron diffraction data. To date, however, too few structures are available for large-scale evaluation purposes (142 structures are currently deposited in the Protein Data Bank (PDB)[70]). Thus, protein structures determined by X-ray crystallography in combination with their electron density data display a foundation for data set assembly.

Here we present a method for placing water molecules on the basis of interaction geometries derived from a large-scale analysis of PDB structures.[71] Our **wa**ter **p**lacement **p**rocedure, WarPP, relies on structural characteristics of protein complexes. Previously defined interaction surfaces[71] have been exploited to define areas that are energetically favorable for water molecules. For validation, we applied WarPP to a high-resolution PDB subset[72] including water molecules with experimental evidence using the electron density of individual atoms (EDIA) score.[72] Last, our placement strategy was applied to single test cases of relevant drug targets:[73] bromodomain and Bruton's tyrosine kinase.

■ METHODS

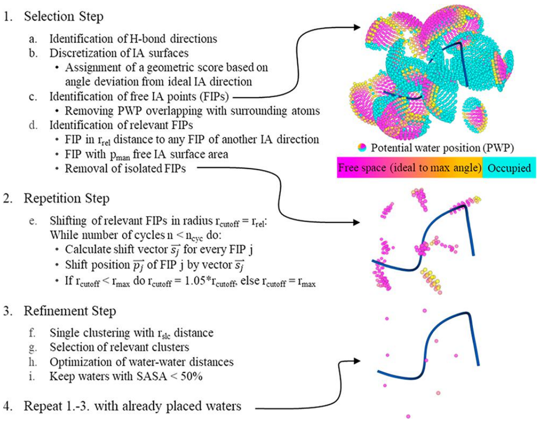
Water Placement. Our method to place water molecules consists of three main steps (Figure 1): (1) identification of free space, (2) generation of explicit water molecules, and (3) refinement of water positions. As a preprocessing step, all of the crystallographically observed water molecules were removed from the structure, and the hydrogen-bond network was optimized using Protoss.[74] The orientations of hydrogen atoms and electron lone pairs are mandatory for WarPP. However, including crystallographically modeled water molecules during the hydrogen-bond network optimization would lead to a substantial bias, with hydrogen atoms and electron lone pairs oriented toward the water molecules. Thus, the crystallographic water molecules were removed, and the dry hydrogen-bond network was optimized to avoid this bias.

Selection Step. Initially, volumes sufficiently large to accommodate a water molecule must be identified (section 1 in Figure 1). Every polar nitrogen or oxygen atom with either an unsaturated hydrogen-bond function or a nonideal interaction geometry was selected. The geometry score from HYDE was applied with a cutoff value of 0.85. This value indicates that a hydrogen bond is statistically less favorable than solvating water molecules.[75] The geometric score is defined by four measurements (Figure 2) and the following combination:

$$f_{\text{dev}}(\text{H-bond}) = f_{\text{dev}}(d_1) \cdot f_{\text{dev}}(d_2) \cdot f_{\text{dev}}(\alpha) \cdot f_{\text{dev}}(\beta) \quad (1)$$

The four geometric measures were selected to symmetrically represent a hydrogen-bonding interaction: the heavy-atom distance (d_1), the distance between the lone pair and the hydrogen atom (the so-called head−head distance, d_2), the donor angle (α), and the acceptor angle (β). Every geometric measure can achieve a score between 0 and 1, depending on the deviation from the ideal distance/angle. Overall, f_{dev}(H-bond) results in a score between 0 and 1.

The polar atoms were assigned interaction directions on the basis of large-scale analysis of hydrogen-bond geometries in protein−ligand complexes[71] using *NAOMI*nova.[76] On the basis of this analysis of hydrogen bonds, interaction surfaces (IASs) were defined.[71] Those surfaces were discretized using a dot distance (d_D) of 0.4 Å (section S1 and Figure S1), with every point resembling a potential water position (PWP). Every PWP was assigned a geometric score according to its position.



1. Selection Step
   a. Identification of H-bond directions
   b. Discretization of IA surfaces
      • Assignment of a geometric score based on angle deviation from ideal IA direction
   c. Identification of free IA points (FIPs)
      • Removing PWP overlapping with surrounding atoms
   d. Identification of relevant FIPs
      • FIP in r_{rel} distance to any FIP of another IA direction
      • FIP with p_{man} free IA surface area
      • Removal of isolated FIPs
2. Repetition Step
   e. Shifting of relevant FIPs in radius $r_{cutoff} = r_{rel}$:
      While number of cycles $n < n_{cyc}$ do:
      • Calculate shift vector $\overrightarrow{s_j}$ for every FIP j
      • Shift position $\overrightarrow{p_j}$ of FIP j by vector $\overrightarrow{s_j}$
      • If $r_{cutoff} < r_{max}$ do $r_{cutoff} = 1.05*r_{cutoff}$, else $r_{cutoff} = r_{max}$
3. Refinement Step
   f. Single clustering with r_{slc} distance
   g. Selection of relevant clusters
   h. Optimization of water-water distances
   i. Keep waters with SASA < 50%
4. Repeat 1.-3. with already placed waters

Potential water position (PWP)
Free space (ideal to max angle)
Occupied

Figure 1. Major steps of the developed water placement workflow, including the main aspects performed in each step (left) and their corresponding effects on the placed water molecules (right); Abbreviations: IA, interaction; PWP, potential water position; FIP, free interaction point; SASA, solvent accessible surface area. Detailed information about the parametrization (parameters: radius for the selection of relevant FIPs, r_{rel}; available interaction surface cutoff for mandatory FIPs, p_{man}; shifting radius cutoff, r_{cutoff}; number of shifting repetitions, n_{cyc}; maximum distance radius, r_{max}; and single linkage clustering radius, r_{slc}) can be found in section S1 in the Supporting Information. The remaining parameters (the shifting vector $\mathbf{s}_j$ and the FIP position $\mathbf{p}_j$) are explained in Methods.

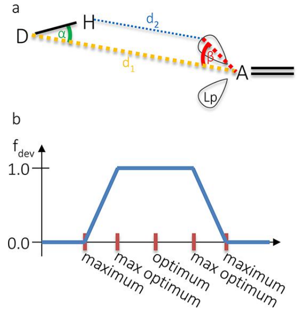


Figure 2. (a) Four measurements for the geometric score calculation. D is the donor heavy atom; H is the hydrogen atom; A is the acceptor heavy atom; Lp is a lone pair; d_1 is the distance between the donor and acceptor heavy atoms; d_2 is the distance between the hydrogen atom and the lone pair, with distances D−H = 1 Å and A−Lp = 1 Å; α is the donor angle (H−D−A); β is the acceptor angle (Lp−A−D). (b) Example plot of the geometric score (f_{dev}) distribution over the angle deviation from the ideal geometry to the maximum allowed deviations. The hydrogen-bond geometry definitions for water molecules are given in Table S1.

The geometric score (f_{dev}) is defined as shown in eq 1 and Figure 2. Here d_1, d_2, and the angle at the water molecule (α for water as the donor, β for water as the acceptor) were set to have ideal geometries. Thus, the score was determined only by the angle at the interacting protein or ligand atom.

The PWPs were then classified as either available (free interaction points (FIPs)) or covered (i.e., other protein or ligand atoms made this position unavailable for a water molecule) (section 1 in Figure 1). For PWP−ligand/protein atom overlap, adjusted atom radii were used to allow hydrogen-bond distances shorter than the sum of the van der Waals radii of polar atoms (see Parametrization of WarPP). All of the remaining, non-overlapping PWPs (i.e., the FIPs) were subsequently used to generate explicit water positions. In theory, a water molecule could be placed on each FIP. However, this would lead to too many water molecules in close proximity. Therefore, relevant FIPs were selected using a distance cutoff of r_{rel} to any surrounding FIPs from other nearby IASs. If no other IASs were within the distance cutoff, the FIP was discarded. The only exceptions were FIPs from "mandatory" IASs with a specific proportion of available surface area (p_{man}). Those FIPs were termed "mandatory" because they often reflect water molecules in narrow cavities. Additionally, isolated FIPs were removed. Since the selection process depends on the surrounding IASs, the removal of FIPs is order-independent.

Repetition Step. Placing a water molecule for every relevant FIP would still result in too many placed waters. Therefore, the following selection strategy was applied (section 2 in Figure 1). In order to detect FIPs that allow multiple hydrogen bonds, the FIPs were shifted toward each other for a specific number of cycles (n_{cyc}). The aim of the iterative shifting procedure was to draw FIPs toward better geometries. The actual shift of each FIP was defined by its own assigned geometric score as well as the geometric scores of the FIPs within a cutoff distance r_{cutoff} (Figure 1).

Every FIP position $\mathbf{p}_j$ is shifted by a shifting vector $\mathbf{s}_j$, which is calculated using all FIPs i within the distance r_{cutoff} and weighted on the basis of the geometric score f_j of the FIP. For every FIP j, geometry-weighted shifting vectors $\mathbf{d}_{ji}$ with the surrounding FIPs i in the distance $\mathrm{r}_{\text{cutoff}}$ are calculated:

$$\mathbf{d}_{ji} = \frac{f_i}{f_i + f_j}\cdot(\mathbf{p}_j - \mathbf{p}_i) \tag{2}$$

For all FIPs i on IAS k (FIP(k)), a normalized shifting vector is calculated:

$$\mathbf{sia}_j(k,\, t) = \frac{\sum_{i\in \text{FIP}(k)} \mathbf{d}_{ji}\cdot f_i}{\sum_{i\in \text{FIP}(k)} f_i} \tag{3}$$

The weighted shifting vector is calculated for every IAS k separately. Then the shifting vector is calculated as

$$\mathbf{s}_j(t) = \frac{\sum_{k\in \text{IAS}} (\mathbf{sia}_j(k,\, t))}{|\text{IAS}|} \tag{4}$$

Because of the normalization by the number of interaction surfaces (|IAS|), every IA contributes equally independent of the actual number of FIPs on each surface. To obtain the final position of FIP j in the current cycle, $p_j(t)$, the shifting vector is added to the position from the previous cycle, $p_j(t-1)$:

$$\mathbf{p}_j(t) = \mathbf{p}_j(t-1) + \mathbf{s}_j(\mathrm{t}) \tag{5}$$

The shifting cycles were repeated to finally result in clusters of FIPs. In every cycle, the distance cutoff r_{cutoff} was increased by a specific percentage up to a final distance cutoff $r_{\max}$.

Refinement of Water Positions. After sufficient convergence of the FIPs was reached, single-linkage clustering was applied with a distance cutoff of r_{slc} to retrieve explicit water positions. The clustering of the points was again weighted by the assigned geometric scores (eq 1). Since only structurally relevant water molecules were of interest, only those positions with less than 50% solvent-accessible surface area (SASA) were kept (50% SASA is about the available surface area when two hydrogen-bonding interactions are formed; data not shown). The placed water molecules (section 2 in Figure 1) are sometimes located in close proximity to each other. Therefore, a water refinement step was added to move water molecules to a correct hydrogen bond distance between them (section 3 in Figure 1).

All of the placed water molecules were optimized using a gradient-based numerical optimization. The scoring function was designed to be simultaneously easily optimizable—continuously differentiable and without singularities (like, e.g., the Lennard-Jones potential has)—and to model the experimentally observable characteristics as closely as possible.

Herein, three different distance terms were modeled: (1) between water molecules and apolar atoms to avoid clashes, (2) between water molecules and polar atoms to represent hydrogen-bonding interactions, and (3) between water molecules to shift them to correct hydrogen-bond distances. The scoring function was the weighted sum of these terms.

The numerical optimization itself was performed using an in-house implementation of the Broyden−Fletcher−Goldfarb−Shanno (BFGS) algorithm (see ref 77 for a detailed description of this algorithm). A simple backtracking line search (described in ref 78) was used to determine the step size. To preserve the positive-definiteness of the Hessian approximation, a damped BFGS update strategy was employed (as described in ref 79). Excessive movement of water molecules during optimization was prevented by limiting the step size such that no water molecule moved more than 0.5 Å in each BFGS iteration.

The whole procedure, from the identification of free space to the optimization of placed water positions, was repeated to account for water networks within the protein structure. Herein, the already-placed water molecules from the previous round were considered in the same way as protein and ligand atoms. For the correct representation of the interaction network, the hydrogen-bond network of the protein−ligand complex, including the placed water molecules, was optimized using Protoss before the second iteration.

Data Sets. *High-Resolution PDB Subset.*[72] To evaluate the developed water placement procedure, WarPP, a previously compiled PDB subset[72] exclusively containing structures with resolution less than 1.5 Å was employed. All of the water molecules contained in this subset were filtered by their underlying electron density using EDIA (preliminary version for water molecules).[72] Furthermore, the high-resolution water molecules were differentiated according to their position as at the protein complex surface, at the protein−ligand interface (PLI), at the protein−protein interface (PPI), or captured (i.e., surrounded by protein). More details about the EDIA calculation and the classification of water molecules are provided in our previous publication.[72] In this study, only the PLI water molecules of relevant ligands (i.e., no buffer

DOI: 10.1021/acs.jcim.8b00271

molecules or common cofactors) were used for evaluation purposes. Because of the recent development of EDIA applicable to multiple atoms ($\mathrm{EDIA_m}$[80]), the ligands were further checked for their quality using a minimum $\mathrm{EDIA_m}$ of 0.8, i.e., a sufficient coverage of electron density. Thus, 1491 protein complexes with 2397 active sites and almost 20 000 water molecules were finally selected (see section S2 for further information).

Drug Targets.[73] Two different protein targets were selected to show the benefit of our method: bromodomain (BRD)[81] and Bruton's tyrosine kinase (BTK).[82−85] Nine BRD and nine BTK structures were selected to analyze the effect of displaced water molecules. Three different aspects were evaluated concerning the quality of our method: recreation of the water position, number of placed water molecules in a specified "area of interest", and recreation of the hydrogen-bond network. Another reason for the selection of these structures was a recent comparison of 3D-RISM, SZMAP, WaterFLAP, and WaterMap.[73] For both targets, structural pairs are available where minor changes to the ligands cause a disruption of the surrounding water network. Thus, those tools were used to place and score water positions. Here we used the first part of the results, the prediction of water positions, to compare our method to state-of-the-art software solutions.

■ RESULTS AND DISCUSSION

This section is split into three parts. First, the parametrization of the water placement procedure, WarPP, is briefly summarized. Second, the large-scale evaluation of WarPP is described. Third, the single test cases for placement of water molecules in the BTK and BRD structures are discussed.

Parametrization of WarPP. Because of the relatively small number of free parameters, they were adjusted manually. The definitions and parametrization of the free parameters of the above-described method for placing water molecules—dot distance (d_D), atom radii adjustment (r_{adj}), radius for relevant FIPs (r_{rel}), proportion of free IA surface area for mandatory FIPs (p_{man}), radius increase (r_i) and maximum radius (r_{max}) for shifting of FIPs, number of cycles (n_{cyc}), and single linkage clustering distance (r_{slc})—are explained in section S1 in the Supporting Information.

Large-Scale Water Placement Evaluation. On the basis of the high-resolution PDB subset,[72] the sensitivity (eq 6) and precision (eq 7) of WarPP were evaluated:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{no. of crystal waters}} \qquad (6)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\text{no. of predicted waters}} \qquad (7)$$

where TP is the number of true positives, which is the number of placed water molecules within X Å of a crystallographically determined water molecule; FP is the number of false positives, given by the number of placed water molecules without a crystallographically determined water molecule within X Å; and FN is the number of false negatives, which is the number of crystallographically determined water molecules without a placed water molecule within X Å. (The number of true negatives (TN), corresponding to "free space" with no crystallographically determined water molecule within X Å, cannot be calculated.)

The active site of the protein was defined as all atoms within a distance r = 8.0 Å of any ligand atom, with the full amino acid side chain considered as soon as one of its atoms is within range. To account for boundary effects (as exemplarily shown in Figure 3), the active-site radius was extended to 10.5 Å for placement of water molecules.

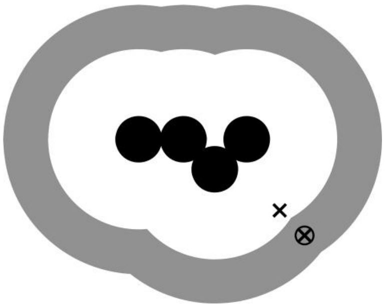


Figure 3. Exemplary representation of active-site selection. Black spheres represent the ligand. The white region is the active site with r = 8.0 Å, and the gray region is the bigger active site with r = 10.5 Å. x denotes a crystallographic water molecule within the active site, and ⊗ indicates a placed water molecule matching the crystallographically observed one.

For sensitivity analysis, only those crystallographic water molecules within 8.0 Å of a ligand atom that had two or more possible hydrogen bonds to either the protein or the ligand were used (19 808 water molecules). The sensitivity gives information about how many crystallographically observed water molecules are not detected by our method. Thus, all of the water molecules interacting with the protein or ligand should be matched by our method.

The method's precision is more difficult to evaluate because the number of water molecules where no crystallographic water molecules are observed is analyzed. However, multiple reasons for nonavailable crystallographic waters are at hand: First, the water might be too flexible to be defined to a specific position, which is frequent for waters at the protein surface and the rim of the active site. Second, water molecules might not have been placed during structure refinement. Our method not only identifies water molecules in confined volumes but also places water molecules that are closer to the surface of the PLI resulting from two or more close interaction surfaces. Therefore, the precision was calculated using all of the crystallographically observed water molecules with sufficient EDIA (0.24 < EDIA < 3.3). Additionally, for the aforementioned reasons, we considered only the placed water molecules classified as at the PLI, at the PPI, or captured.

Overall, our method can recall 48% of the crystallographically observed water molecules within 0.5 Å and 80% within 1.0 Å (Figure 4a). The precision of our method is 66% at 1.0 Å (Figure 4b), which means that on average our method places three water molecules for two crystallographically observed ones. However, as mentioned before, some of the water molecules might not be resolved by crystallography for various reasons. Therefore, in addition to the sensitivity and precision, we analyzed the pairwise distance and EDIA distribution of the placed water molecules. In the case that many water molecules are closely placed, the interpretation of the results is very difficult. The pairwise distance distribution of

DOI: 10.1021/acs.jcim.8b00271

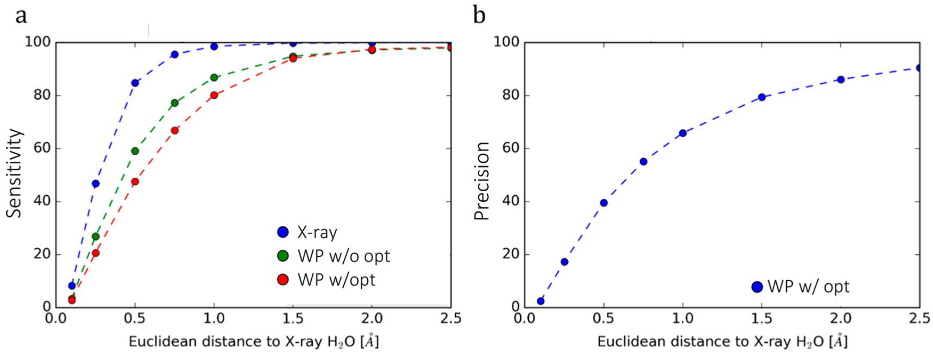


Figure 4. (a) Sensitivity of optimization of crystallographically observed water molecules (X-ray) and the water placement procedure with (WP w/opt) and without optimization (WP w/o opt). The optimization of crystallographically observable water molecules should result in only minor movements, which indicates a correct parametrization of the optimization function. (b) Precision of WarPP with optimization.

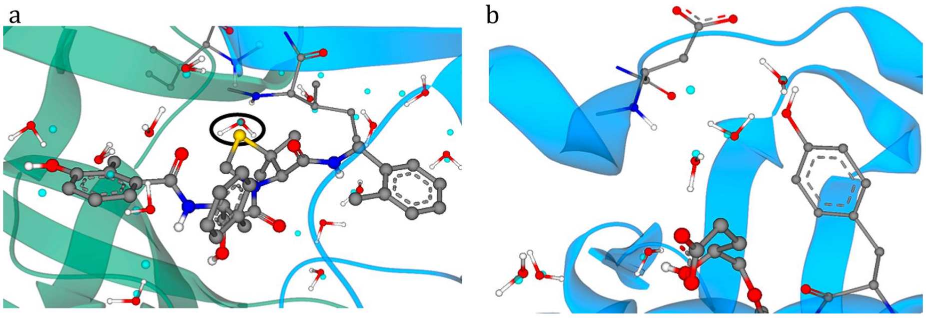


Figure 5. Example cases of water placement. (a) Accurate placement of a highly integrated water molecule in HIV protease (PDB ID 1kzk,[86] HOH-A-1037, circled in black). (b) Recreation of a water network (PDB ID 3az3[87]). Light-blue spheres represent placed water molecules; water molecules shown as a red oxygen atom with white hydrogen atoms are X-ray-observed water molecules.

crystallographically observed water molecules showed the main peak at around 2.8 Å, the ideal hydrogen-bond distance (Figure S11). The distribution of the placed water molecules is in accordance with the distribution of the crystallographically observed water molecules (Figure S12). Thus, it also demonstrates that the parametrization of the optimized scoring function accurately captures the experimental observations. The EDIA distributions were generated for crystallographically observed water molecules, water molecules placed by WarPP, and randomly placed water molecules using a grid-based approach (Figure S13). For randomly placed water molecules, a 3D grid was generated for the active site with a voxel distance of 2.2 Å. Every available (i.e., not covered by protein or ligand atoms) voxel was used for calculating the EDIA. In order to achieve better sampling, the 3D grid was shifted separately in the *x*, *y*, and *z* directions for half the voxel distance. The distribution of EDIA values for water molecules placed by WarPP has a greater number of small (i.e., bad) EDIA values compared with the EDIA distribution of crystallographically observed water molecules. However, the EDIA distribution of placed water molecules is about 40% better than that of randomly placed water molecules.

An overview of our water placement shows that many interface water molecules were matched well (Figure 5a). Also, the tetrahedrally coordinated water molecule in HIV protease (HOH-A-1037) was placed with great accuracy (0.14 Å). Water networks are also recreated in our water placement procedure (Figure 5b). Further examples of the consistency of WarPP are shown in Figure 6. In addition to crystallographically observable water molecules, WarPP also predicts water positions where electron density is available but no water molecules were placed (Figure 6a). In this example, the unaccounted electron density was within the active site, and the placed water molecule was within hydrogen-bond distance to a backbone nitrogen and could potentially interact with another water molecule. An example of the limitations of WarPP, provided by the thrombin S1 pocket, is shown in Figure 6c. The deeply buried water molecule (HOH-H-3098) was not matched by a placed water molecule. The reason is the identification of free interaction directions. The only possible interaction for this particular water molecule was the backbone carbonyl oxygen of Phe-H-227. However, this oxygen atom already forms two geometrically high-quality interactions to two backbone nitrogen atoms (Ser-H-214 and Trp-H-215). Thus, no free interaction direction is available for which a water molecule would need to be placed, even though the free-space identification correctly identified a small number of available points around the crystallographic water molecule.

Single-Case Evaluation. On the basis of three different protein families that were previously used for the evaluation of water placement and scoring methods,[73] the accuracy of water molecule placement was evaluated as well as compared to that by other state-of-the-art methods: 3D-RISM, SZMAP, WaterFLAP, and WaterMap.

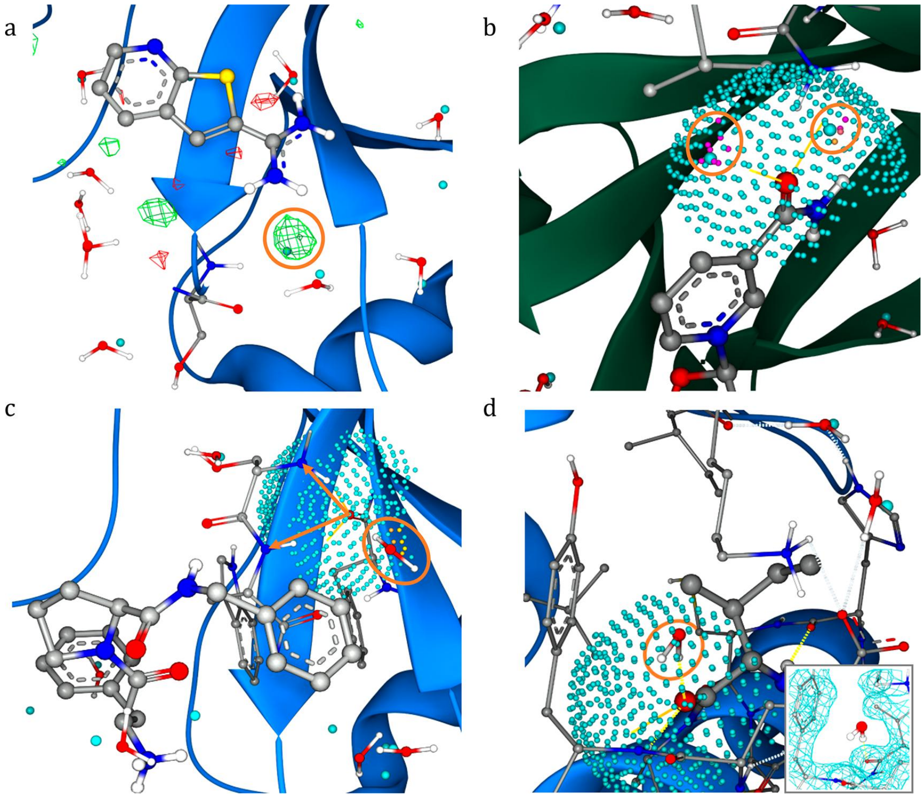


Figure 6. Example cases showing the consistency of WarPP. (a) No X-ray water, electron density, placed water: water molecules are correctly placed in unaccounted electron density (green mesh) (PDB ID 1c5u[88]). (b) No X-ray water, no electron density, placed water: water molecules are placed in confined volumes (PDB ID 1j96[89]). (c) X-ray water, electron density, no placed water: an unidentified water molecule (yellow circle, HOH-H-3098) is in a tight cavity of thrombin; yellow arrows indicate hydrogen-bonding interactions of the carbonyl oxygen of Phe-H-3098 to the backbone nitrogen atoms of Ser-H-214 and Trp-H215 (PDB ID 2zff[90]). (d) X-ray water, no electron density, no placed water: there is a narrow polar area within the protein with no free space; the crystallographically placed water molecule has no electron density (inset) (PDB ID 3fpc[91]). Light blue spheres represent placed water molecules; water molecules shown as a red oxygen atom with white hydrogen atoms are X-ray-observed water molecules.

The placement of water molecules for different protein targets was compared to the placement in the previous study.[73] Herein, the placement was evaluated concerning three different aspects: (1) the distance of each crystallographically observed water molecule to the closest placed water molecule, with every placed water molecule considered only once; (2) the number of placed water molecules in a certain area of interest, i.e., to evaluate the ease of interpretation of the method; and (3) the recreation of the water network.

The distances between the oxygen atoms of the observed and placed water molecules were measured to quantify the accuracy of the placement procedure. Our water placement procedure placed 81% of water molecules in BRD structures (Figure 7a,b) and 60% of those in BTK structures (Figure 7c,d) within 1.0 Å of the crystallographically observed ones. In the previous study, WaterMap achieved the highest accuracies for both targets (~90%). The accuracies for BRD varied for the other tools between 60% (3D-RISM) and 78% (WaterFLAP), while apart from WaterMap, all of the other tools achieved accuracies of around 60% for BTK.

Apart from the single distances, the number of placed water molecules gives valuable information about the interpretability of the results. Therefore, a so-called "area of interest" was defined using spheres that resembled the active site. The number of placed water molecules within this area was counted for each software tool and compared to the number of crystallographically observed ones. WarPP placed the same number of water molecules as crystallographically observable ones in both targets. Here one must point out that the total number of water molecules placed for each structure individually did not always match exactly. WaterMap also placed roughly the same number of water molecules as observed experimentally, while SZMAP and WaterFLAP placed 30−60% more water molecules. The numbers of water molecules placed by 3D-RISM were fewer in BRD structures and more in BTK structures compared with the numbers of crystallographically observable waters.

Last, the recreation of the water network was analyzed using pairwise distances between the oxygen atoms of the water molecules. The pairwise distances were calculated for the crystallographically observed water molecules and separately

DOI: 10.1021/acs.jcim.8b00271

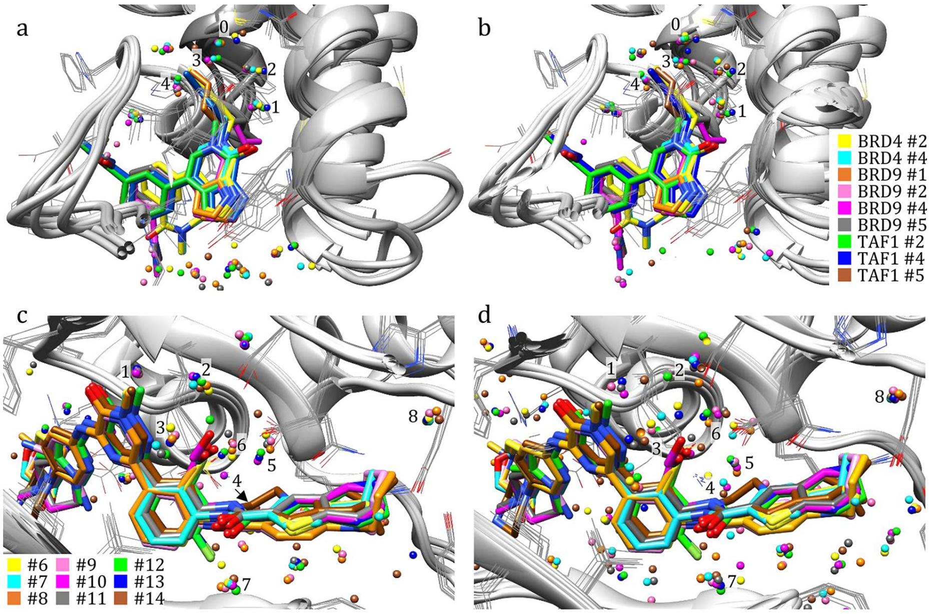


Figure 7. Superimposed structures of (a) BRD with crystallographically observed water molecules, (b) BRD with water molecules placed by WarPP, (c) BTK with crystallographically observed water molecules, and (d) BTK with water molecules placed by WarPP. A list of corresponding PDB IDs can be found in Table S3. Molecular graphics were generated using Chimera.[92]

for the placed water molecules. As an objective criterion, the networks were then compared using root-mean-square deviation (RMSD) values. The water network is important especially for further usage of the water placement (i.e., scoring). The water molecules might be close to each other and thus form hydrogen bonds. If the water placement cannot correctly recreate those distances, the formation of hydrogen bonds might not be possible. Compared with the state-of-the-art methods, WarPP was best in recreating the pairwise distances in BRD structures (average RMSD = 0.51 Å) and second best for BTK (average RMSD = 0.69 Å), while WaterMap was most accurate for BTK structures (average RMSD = 0.43 Å).

Overall, our water placement procedure based on geometric criteria derived from protein crystal structures, WarPP, gave results comparable to those of state-of-the-art software solutions based on the selected 18 structures of BRD and BTK targets. In addition to the achieved sensitivity, WarPP is fast, with an average run time of 6.8 s per active site (Figure 8). It is thus well able not only to place water positions for docking poses but also to generate different water networks upon exploitation of protein flexibility.

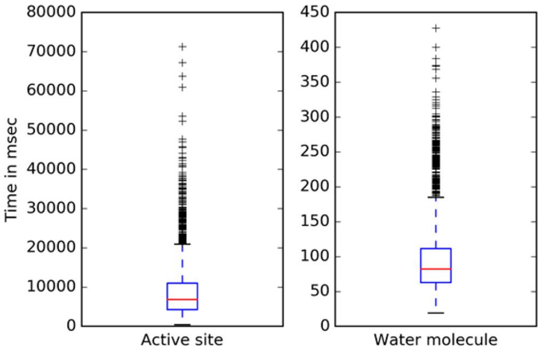


Figure 8. Run times for placing water molecules within an active site of 10.5 Å (left) and for each water molecule placed (right).

■ CONCLUSIONS

Many methods for the evaluation of water molecules have been developed throughout the last years. One great drawback of nearly every method is the evaluation strategy, which is often based on a few selected protein structures only. Additionally, a major problem is the lack of experimental data for single water molecules. Individual energy contributions cannot be measured experimentally, and the only experimental data available in substantial quantities to support the position of water molecules is electron density. Therefore, we used the previously developed EDIA score[72,80] as a basis for our evaluation of water molecules well-supported by electron density.

Here we have described a new method for placing water molecules in protein structures. On the basis of a large-scale evaluation of 2400 protein−ligand interfaces containing 20 000 water molecules, this evaluation is to our knowledge the most extensive among water placement methods. Our geometry-based method achieved a sensitivity of 80% for placement of a water molecule within 1.0 Å of a crystallographically observed water molecule.

A comparison to state-of-the-art methods based on two relevant drug targets with nine protein−ligand structures each showed that our developed water placement procedure is at least comparable and in some cases even superior to those methods.

WarPP was designed to place water molecules in protein−ligand structures resolved by X-ray crystallography. For the high-resolution data set, structures containing DNA or RNA were excluded. In direct connection to the mentioned aspect is the composition of the evaluation data set. A data set for which multiple lines of evidence exist (i.e., neutron diffraction data) would be ideal. However, in our opinion such idealistic data sets that are additionally freely available to the public do not exist for many structures. Additionally, we think that using an idealistic but small data set for a computational method such as the presented one would not allow a statistical valid evaluation. Thus, using a large data set, such as our high-resolution data set, allows the determination of a realistic application range. On the basis of the free hydrogen-bond directions as a starting point for WarPP, water molecules that are at least partially enthalpically stabilized are placed. From our point of view, a detailed reason for the stabilization of water molecules in protein structures cannot always be identified. However, water molecules that are solely entropically stabilized, in so-called "hydrophobic bubbles" completely surrounded by hydrophobic residues, cannot be placed with our method. Since our method relies on freely available hydrogen-bond functions, the latter water molecules cannot be predicted. According to one of our previous publications,[72] among all water molecules with experimental evidence in protein crystal structures, about 0.5% are in hydrophobic bubbles. An often-mentioned problem of water placement methods is the recreation of water−water networks. This problem is addressed in part by WarPP's two iteration processes. In the second iteration, water molecules placed in the first one are considered, and water−water contacts are thus modeled. However, this does not lead to a "full" solvation shell of the protein complex.

The investigation of water placement as well as the contribution of individual water scores, such as recently published by Aldeghi and co-workers,[93] can aid a better understanding of the role of water molecules in protein structures and help during drug development strategies.

Finally, the water placement procedure based on geometric criteria is reproducible and comprehensible in combination with a short run time. WarPP allows consistent placement in protein−ligand structures where no water molecules were placed (i.e., due to low resolution of the crystal structure) as well as for docking poses. Furthermore, it displays a solid foundation for further analysis of protein−ligand complexes as well as scoring of protein−ligand complexes with a consistent and reliable representation of water molecules.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00271.

Parametrization and supporting figures and tables (PDF)

Comma-separated table with PDB IDs, molecule IDs, and water IDs (TXT)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

ORCID

Eva Nittinger: 0000-0001-7231-7996

Florian Flachsenberg: 0000-0001-7051-8719

Matthias Rarey: 0000-0002-9553-6531

Funding

This project was part of the Biokatalyse2021 Cluster and was funded by the Federal Ministry of Education and Research (BMBF) under Grant 031A183B.

Notes

The authors declare no competing financial interest.

A web server for the placement of water molecules is in preparation and will be integrated into the ProteinsPlus web service, available at https://proteins.plus. Additionally, WarPP will be made available as part of the NAOMI ChemBio Suite as stand-alone tool, available at https://software.zbh.uni-hamburg.de/, and will be free for academic use and evaluation purposes. Structural preprocessing functionality, including Protoss and EDIA, is also available as a web service at http://proteins.plus.

■ ACKNOWLEDGMENTS

We thank Daniel F. Ortwine and Paul Gibbons from Genentech for providing the BRD and BTK structures and for fruitful discussions.

■ ABBREVIATIONS

PWP, potential water position; FIP, free interaction point; EDIA, electron density of individual atoms; BRD, bromodomain; BTK, Bruton's tyrosine kinase.

■ REFERENCES

(1) Breiten, B.; Lockett, M. R.; Sherman, W.; Fujita, S.; Al-Sayah, M.; Lange, H.; Bowers, C. M.; Heroux, A.; Krilov, G.; Whitesides, G. M. Water Networks Contribute to Enthalpy/Entropy Compensation in Protein-Ligand Binding. *J. Am. Chem. Soc.* **2013**, *135* (41), 15579−15584.

(2) Rühmann, E.; Betz, M.; Heine, A.; Klebe, G. Fragment Binding Can Be Either More Enthalpy-Driven or Entropy-Driven: Crystal Structures and Residual Hydration Patterns Suggest Why. *J. Med. Chem.* **2015**, *58* (17), 6960−6971.

(3) 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.

(4) 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.

(5) 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.

(6) 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.

(7) Kellogg, G. E.; Chen, D. L. The Importance of Being Exhaustive. Optimization of Bridging Structural Water Molecules and Water Networks in Models of Biological Systems. *Chem. Biodiversity* **2004**, *1* (1), 98−105.

(8) 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.

(9) Pitt, W. R.; Goodfellow, J. M. Modelling of Solvent Positions around Polar Groups in Proteins. *Protein Eng., Des. Sel.* **1991**, *4* (5), 531−537.

(10) Pitt, W. R.; Murray-Rust, J.; Goodfellow, J. M. AQUARIUS2: Knowledge-Based Modeling of Solvent Sites around Proteins. *J. Comput. Chem.* **1993**, *14* (9), 1007−1018.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) Schymkowitz, J. W. H.; Rousseau, F.; Martins, I. C.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L. Prediction of Water and Metal Binding Sites and Their Affinities by Using the Fold-X Force Field. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (29), 10147−10152.

(16) 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.

(17) Morozenko, A.; Stuchebrukhov, A. A. Dowser++, a New Method of Hydrating Protein Structures. *Proteins: Struct., Funct., Genet.* **2016**, *84* (10), 1347−1357.

(18) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849−857.

(19) Wade, R.; Clark, K.; Goodford, P. Further Development of Hydrogen Bond Functions for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. 1. Ligand Probe Groups with the Ability to Form Two Hydrogen Bonds. *J. Med. Chem.* **1993**, *36* (1), 140−147.

(20) Wade, R. C.; Goodford, P. J. Further Development of Hydrogen Bond Functions for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. 2. Ligand Probe Groups with the Ability To Form More Than Two Hydrogen Bonds. *J. Med. Chem.* **1993**, *36* (1), 148−156.

(21) Kovalenko, A.; Hirata, F. Self-Consistent Description of a Metal−Water Interface by the Kohn−Sham Density Functional Theory and the Three-Dimensional Reference Interaction Site Model. *J. Chem. Phys.* **1999**, *110* (20), 10095−10112.

(22) 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.

(23) Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. *Proteins: Struct., Funct., Genet.* **1991**, *11* (1), 29−34.

(24) Evensen, E.; Joseph-McCarthy, D.; Weiss, G. A.; Schreiber, S. L.; Karplus, M. Ligand Design by a Combinatorial Approach Based on Modeling and Experiment: Application to HLA-DR4. *J. Comput.-Aided Mol. Des.* **2007**, *21* (7), 395−418.

(25) Bitetti-Putzer, R.; Joseph-McCarthy, D.; Hogle, J. M.; Karplus, M. Functional Group Placement in Protein Binding Sites: A Comparison of GRID and MCSS. *J. Comput.-Aided Mol. Des.* **2001**, *15* (10), 935−960.

(26) 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.

(27) Zheng, M.; Li, Y.; Xiong, B.; Jiang, H.; Shen, J. Water PMF for Predicting the Properties of Water Molecules in Protein Binding Site. *J. Comput. Chem.* **2013**, *34* (7), 583−592.

(28) 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.

(29) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein−ligand Binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (3), 808−813.

(30) 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.

(31) 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), 044101.

(32) Li, Z.; Lazaridis, T. Computing the Thermodynamic Contributions of Interfacial Water. *Methods Mol. Biol.* **2012**, *819*, 393−404.

(33) López, E. D.; Arcon, J. P.; Gauto, D. F.; Petruk, A. A.; Modenutti, C. P.; Dumas, V. G.; Marti, M. A.; Turjanski, A. G. WATCLUST: A Tool for Improving the Design of Drugs Based on Protein-Water Interactions. *Bioinformatics* **2015**, *31* (22), 3697−3699.

(34) 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.

(35) Hu, B.; Lill, M. A. WATsite: Hydration Site Prediction Program with PyMOL Interface. *J. Comput. Chem.* **2014**, *35* (16), 1255−1260.

(36) Yang, Y.; Hu, B.; Lill, M. A. WATsite2.0 with PyMOL Plugin: Hydration Site Prediction and Visualization. *Methods Mol. Biol.* **2017**, *1611*, 123−134.

(37) Gerogiokas, G.; Calabro, G.; Henchman, R. H.; Southey, M. W. Y.; Law, R. J.; Michel, J. Prediction of Small Molecule Hydration Thermodynamics with Grid Cell Theory. *J. Chem. Theory Comput.* **2014**, *10* (1), 35−48.

(38) Michel, J.; Henchman, R. H.; Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Law, R. J. Evaluation of Host-Guest Binding Thermodynamics of Model Cavities with Grid Cell Theory. *J. Chem. Theory Comput.* **2014**, *10* (9), 4055−4068.

(39) Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Hefeitz, A.; Bodkin, M.; Law, R. J.; Michel, J. Evaluation of Water Displacement Energetics in Protein Binding Sites with Grid Cell Theory. *Phys. Chem. Chem. Phys.* **2015**, *17* (13), 8416−8426.

(40) Zia, S. R.; Gaspari, R.; Decherchi, S.; Rocchia, W. Probing Hydration Patterns in Class-A GPCRs via Biased MD: The A2A Receptor. *J. Chem. Theory Comput.* **2016**, *12* (12), 6049−6061.

(41) Woods, C. J.; Essex, J. W.; King, M. A. The Development of Replica-Exchange-Based Free-Energy Methods. *J. Phys. Chem. B* **2003**, *107* (49), 13703−13710.

(42) Woods, C. J.; Essex, J. W.; King, M. A. Enhanced Configurational Sampling in Binding Free-Energy Calculations. *J. Phys. Chem. B* **2003**, *107*, 13711−13718.

(43) Gilson, M. K. K.; Given, J. A. A.; Bush, B. L. L.; McCammon, J. A. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72* (3), 1047−1069.

(44) Hamelberg, D.; McCammon, J. A. Standard Free Energy of Releasing a Localized Water Molecule from the Binding Pockets of Proteins: Double-Decoupling Method. *J. Am. Chem. Soc.* **2004**, *126* (24), 7683−7689.

(45) 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.

(46) 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), 19.

(47) 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.

(48) Adams, D. J. Chemical Potential of Hard-Sphere Fluids by Monte Carlo Methods. *Mol. Phys.* **1974**, *28* (5), 1241−1252.

(49) Adams, D. J. Grand Canonical Ensemble Monte Carlo for a Lennard-Jones Fluid. *Mol. Phys.* **1975**, *29* (1), 307−311.

DOI: 10.1021/acs.jcim.8b00271

(50) Woo, H.-J.; Dinner, A. R.; Roux, B. Grand Canonical Monte Carlo Simulations of Water in Protein Environments. *J. Chem. Phys.* **2004**, *121* (13), 6392−6400.

(51) Biedermannová, L.; Schneider, B. Hydration of Proteins and Nucleic Acids: Advances in Experiment and Theory. A Review. *Biochim. Biophys. Acta, Gen. Subj.* **2016**, *1860* (9), 1821−1835.

(52) 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.

(53) Carugo, O. Correlation between Occupancy and B Factor of Water Molecules in Protein Crystal Structures. *Protein Eng., Des. Sel.* **1999**, *12* (12), 1021−1024.

(54) Kellogg, G. E.; Abraham, D. J. Hydrophobicity: Is LogP$_{o/w}$ More than the Sum of Its Parts? *Eur. J. Med. Chem.* **2000**, *35* (7−8), 651−661.

(55) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58* (3), 380−388.

(56) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455−461.

(57) Morozenko, A.; Leontyev, I. V.; Stuchebrukhov, A. A. Dipole Moment and Binding Energy of Water in Proteins from Crystallographic Analysis. *J. Chem. Theory Comput.* **2014**, *10* (10), 4618−4623.

(58) Wade, R. C. Solvation of the Active Site of Cytochrome P450-Cam. *J. Comput.-Aided Mol. Des.* **1990**, *4* (2), 199−204.

(59) Helms, V.; Wade, R. C. Thermodynamics of Water Mediating Protein-Ligand Interactions in Cytochrome P450cam: A Molecular Dynamics Study. *Biophys. J.* **1995**, *69* (3), 810−824.

(60) Henchman, R. H.; McCammon, J. A. Structural and Dynamic Properties of Water around Acetylcholinesterase. *Protein Sci.* **2002**, *11* (9), 2080−2090.

(61) Olkhova, E.; Hutter, M. C.; Lill, M. a.; Helms, V.; Michel, H. Dynamic Water Networks in Cytochrome C Oxidase from Paracoccus Denitrificans Investigated by Molecular Dynamics Simulations. *Biophys. J.* **2004**, *86* (4), 1873−1889.

(62) Sindhikara, D. J.; Yoshida, N.; Hirata, F. Placevent: An Algorithm for Prediction of Explicit Solvent Atom Distribution-Application to HIV-1 Protease and F-ATP Synthase. *J. Comput. Chem.* **2012**, *33* (18), 1536−1543.

(63) Lazaridis, T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. *J. Phys. Chem. B* **1998**, *102* (18), 3531−3541.

(64) Lazaridis, T. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids. *J. Phys. Chem. B* **1998**, *102* (18), 3542−3550.

(65) Cappel, D.; Sherman, W.; Beuming, T. Calculating Water Thermodynamics in the Binding Site of Proteins—Applications of WaterMap to Drug Discovery. *Curr. Top. Med. Chem.* **2017**, *17* (23), 2586−2598.

(66) Wickstrom, L.; Deng, N.; He, P.; Mentes, A.; Nguyen, C.; Gilson, M. K.; Kurtzman, T.; Gallicchio, E.; Levy, R. M. Parameterization of an Effective Potential for Protein-Ligand Binding from Host-Guest Affinity Data. *J. Mol. Recognit.* **2016**, *29* (1), 10−21.

(67) Ben-Naim, A.; Marcus, Y. Solvation Thermodynamics of Nonionic Solutes. *J. Chem. Phys.* **1984**, *81* (4), 2016−2027.

(68) Ross, G. A.; Bodnarchuk, M. S.; Essex, J. W. Water Sites, Networks, and Free Energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc.* **2015**, *137* (47), 14930−14943.

(69) Blow, D. Electron-Density Maps. In *Outline of Crystallography for Biologists*; Oxford University Press: Oxford, U.K., 2002; p 196.

(70) 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 Res.* **2000**, *28* (1), 235−242.

(71) Nittinger, E.; Inhester, T.; Bietz, S.; Meyder, A.; Schomburg, K. T.; Lange, G.; Klein, R.; Rarey, M. Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein−Ligand Interfaces. *J. Med. Chem.* **2017**, *60* (10), 4245−4257.

(72) Nittinger, E.; Schneider, N.; Lange, G.; Rarey, M. Evidence of Water Molecules—A Statistical Evaluation of Water Molecules Based on Electron Density. *J. Chem. Inf. Model.* **2015**, *55* (4), 771−783.

(73) Nittinger, E.; Gibbons, P.; Eigenbrot, C.; Davies, D. R.; Maurer, B.; Yu, C. L.; Kiefer, J. R.; Kugelstatter, A.; Murray, J.; Ortwine, D. F.; Tang, Y.; Tsui, V. Water Molecules in Protein−Ligand Interfaces. Evaluation of Software Tools and SAR Comparison. Unpublished data, 2018.

(74) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminf.* **2014**, *6* (1), 12.

(75) Schneider, N.; Lange, G.; Hindle, S.; Klein, R.; Rarey, M. A Consistent Description of HYdrogen Bond and DEhydration Energies in Protein-Ligand Complexes: Methods behind the HYDE Scoring Function. *J. Comput.-Aided Mol. Des.* **2013**, *27* (1), 15−29.

(76) Inhester, T.; Nittinger, E.; Sommer, K.; Schmidt, P.; Bietz, S.; Rarey, M. *NAOMI*nova: Interactive Geometric Analysis of Noncovalent Interactions in Macromolecular Structures. *J. Chem. Inf. Model.* **2017**, *57* (9), 2132−2142.

(77) Nocedal, J.; Wright, S. J. Quasi-Newton Methods. In *Numerical Optimization*; Springer Series in Operations Research and Financial Engineering; Springer: New York, 2006; pp 135−163.

(78) Nocedal, J.; Wright, S. J. Line Search Methods. In *Numerical Optimization*; Springer Series in Operations Research and Financial Engineering; Springer: New York, 2006; pp 30−65.

(79) Nocedal, J.; Wright, S. J. Sequential Quadratic Programming. In *Numerical Optimization*; Springer Series in Operations Research and Financial Engineering; Springer: New York, 2006; pp 529−562.

(80) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. *J. Chem. Inf. Model.* **2017**, *57* (10), 2437−2447.

(81) Crawford, T. D.; Tsui, V.; Flynn, E. M.; Wang, S.; Taylor, A. M.; Côté, A.; Audia, J. E.; Beresini, M. H.; Burdick, D. J.; Cummings, R.; Dakin, L. A.; Duplessis, M.; Good, A. C.; Hewitt, M. C.; Huang, H. R.; Jayaram, H.; Kiefer, J. R.; Jiang, Y.; Murray, J.; Nasveschuk, C. G.; Pardo, E.; Poy, F.; Romero, F. A.; Tang, Y.; Wang, J.; Xu, Z.; Zawadzke, L. E.; Zhu, X.; Albrecht, B. K.; Magnuson, S. R.; Bellon, S.; Cochran, A. G. Diving into the Water: Inducible Binding Conformations for BRD4, TAF1(2), BRD9, and CECR2 Bromodomains. *J. Med. Chem.* **2016**, *59* (11), 5391−5402.

(82) Johnson, A. R.; Kohli, P. B.; Katewa, A.; Gogol, E.; Belmont, L. D.; Choy, R.; Penuel, E.; Burton, L.; Eigenbrot, C.; Yu, C.; Ortwine, D. F.; Bowman, K.; Franke, Y.; Tam, C.; Estevez, A.; Mortara, K.; Wu, J.; Li, H.; Lin, M.; Bergeron, P.; Crawford, J. J.; Young, W. B. Battling Btk Mutants with Noncovalent Inhibitors That Overcome Cys481 and Thr474 Mutations. *ACS Chem. Biol.* **2016**, *11* (10), 2897−2907.

(83) Di Paolo, J. A.; Huang, T.; Balazs, M.; Barbosa, J.; Barck, K. H.; Bravo, B. J.; Carano, R. A. D.; Darrow, J.; Davies, D. R.; DeForge, L. E.; Diehl, L.; Ferrando, R.; Gallion, S. L.; Giannetti, A. M.; Gribling, P.; Hurez, V.; Hymowitz, S. G.; Jones, R.; Kropf, J. E.; Lee, W. P.; Maciejewski, P. M.; Mitchell, S. A.; Rong, H.; Staker, B. L.; Whitney, J. A.; Yeh, S.; Young, W. B.; Yu, C.; Zhang, J.; Reif, K.; Currie, K. S. Specific Btk Inhibition Suppresses B Cell− and Myeloid Cell−mediated Arthritis. *Nat. Chem. Biol.* **2011**, *7* (1), 41−50.

(84) Young, W. B.; Barbosa, J.; Blomgren, P.; Bremer, M. C.; Crawford, J. J.; Dambach, D.; Eigenbrot, C.; Gallion, S.; Johnson, A. R.; Kropf, J. E.; Lee, S. H.; Liu, L.; Lubach, J. W.; MacAluso, J.; MacIejewski, P.; Mitchell, S. A.; Ortwine, D. F.; Di Paolo, J.; Reif, K.; Scheerens, H.; Schmitt, A.; Wang, X.; Wong, H.; Xiong, J. M.; Xu, J.; Yu, C.; Zhao, Z.; Currie, K. S. Discovery of Highly Potent and Selective Bruton's Tyrosine Kinase Inhibitors: Pyridazinone Analogs with Improved Metabolic Stability. *Bioorg. Med. Chem. Lett.* **2016**, *26* (2), 575−579.

(85) Young, W. B.; Barbosa, J.; Blomgren, P.; Bremer, M. C.; Crawford, J. J.; Dambach, D.; Gallion, S.; Hymowitz, S. G.; Kropf, J. E.; Lee, S. H.; Liu, L.; Lubach, J. W.; Macaluso, J.; Maciejewski, P.; Maurer, B.; Mitchell, S. A.; Ortwine, D. F.; Di Paolo, J.; Reif, K.; Scheerens, H.; Schmitt, A.; Sowell, C. G.; Wang, X.; Wong, H.; Xiong,

J.-M.; Xu, J.; Zhao, Z.; Currie, K. S. Potent and Selective Bruton's Tyrosine Kinase Inhibitors: Discovery of GDC-0834. *Bioorg. Med. Chem. Lett.* **2015**, *25* (6), 1333−1337.

(86) Reiling, K. K.; Endres, N. F.; Dauber, D. S.; Craik, C. S.; Stroud, R. M. Anisotropic Dynamics of the JE-2147-HIV Protease Complex: Drug Resistance and Thermodynamic Binding Mode Examined in a 1.09 A Structure. *Biochemistry* **2002**, *41* (14), 4582−4594.

(87) Kashiwagi, H.; Ono, Y.; Shimizu, K.; Haneishi, T.; Ito, S.; Iijima, S.; Kobayashi, T.; Ichikawa, F.; Harada, S.; Sato, H.; Sekiguchi, N.; Ishigai, M.; Takahashi, T. Novel Nonsecosteroidal Vitamin D 3 Carboxylic Acid Analogs for Osteoporosis, and SAR Analysis. *Bioorg. Med. Chem.* **2011**, *19* (16), 4721−4729.

(88) Katz, B. A.; Mackman, R.; Luong, C.; Radika, K.; Martelli, A.; Sprengeler, P. A.; Wang, J.; Chan, H.; Wong, L. Structural Basis for Selectivity of a Small Molecule, S1-Binding, Submicromolar Inhibitor of Urokinase-Type Plasminogen Activator. *Chem. Biol.* **2000**, *7* (4), 299−312.

(89) Nahoum, V.; Gangloff, A.; Legrand, P.; Zhu, D. W.; Cantin, L.; Zhorov, B. S.; Luu-The, V.; Labrie, F.; Breton, R.; Lin, S. X. Structure of the Human 3a-Hydroxysteroid Dehydrogenase Type 3 in Complex with Testosterone and NADP at 1.25-A Resolution. *J. Biol. Chem.* **2001**, *276* (45), 42091−42098.

(90) Baum, B.; Mohamed, M.; Zayed, M.; Gerlach, C.; Heine, A.; Hangauer, D.; Klebe, G. More than a Simple Lipophilic Contact: A Detailed Thermodynamic Analysis of Nonbasic Residues in the S1 Pocket of Thrombin. *J. Mol. Biol.* **2009**, *390* (1), 56−69.

(91) Goihberg, E.; Peretz, M.; Tel-Or, S.; Dym, O.; Shimon, L.; Frolow, F.; Burstein, Y. Biochemical and Structural Properties of Chimeras Constructed by Exchange of Cofactor-Binding Domains in Alcohol Dehydrogenases from Thermophilic and Mesophilic Microorganisms. *Biochemistry* **2010**, *49* (9), 1943−1953.

(92) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605−1612.

(93) Aldeghi, M.; Ross, G. A.; Bodkin, M. J.; Essex, J. W.; Knapp, S.; Biggin, P. C. Large-Scale Analysis of Water Stability in Bromodomain Binding Pockets with Grand Canonical Monte Carlo. *Commun. Chem.* **2018**, *1* (1), 19.

DOI: 10.1021/acs.jcim.8b00271

Water Molecules in Protein-Ligand Interfaces. Evaluation of Software Tools and SAR Comparison.

[P2] **Nittinger, E.**; Gibbons, P.; Eigenbrot, C.; Davies, D. R.; Maurer, B.; Yu, C. L.; Kiefer, J. R.; Kuglstatter, A.; Murray, J.; Ortwine, D. F.; Tang, Y.; Tsui, V. Water Molecules in Protein-Ligand Interfaces. Evaluation of Software Tools and SAR Comparison. J. Comput. Aided. Mol. Des., *Submitted for publication*.

Manuscript submitted to Journal of Computer Aided Molecular Design and under review at the time this dissertation was officially deposited at the University of Hamburg.

Water Molecules in Protein-Ligand Interfaces. Evaluation of Software Tools and SAR Comparison

*Eva Nittinger*1, Paul Gibbons*2, Charles Eigenbrot*2, Doug R. Davies3, Brigitte Maurer2, Christine L. Yu2, James R. Kiefer2, Andreas Kuglstatter4, Jeremy Murray2, Daniel F. Ortwine2, Yong Tang5,††, Vickie Tsui2,†*

1: Universität Hamburg, ZBH – Center for Bioinformatics, 20251 Hamburg, Germany
2: Genentech, 1DNA Way, South San Francisco, CA 94080 USA
3: Beryllium Discovery, 7869 NE Day Road West, Bainbridge Island, WA 98110 USA
4: F. Hoffman-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland
5: Constellation Pharmaceuticals, 215 First Street, Cambridge, MA 02142 USA

ORCID

Eva Nittinger 0000-0001-7231-7996

Corresponding Authors

* Paul Gibbons: gibbons.paul@gene.com
* Eva Nittinger: nittinger@zbh.uni-hamburg.de
* Charles Eigenbrot: charleseigenbrot@comcast.net

KEYWORDS. Water; Water prediction; Water placement; Water scoring; 3D-RISM; SZMAP; WaterFLAP; WaterMap; WaterRank. BTK; Bruton's kinase; BRD; bromodomain

Abstract. Targeting the interaction with or displacement of the 'right' water molecule can significantly increase inhibitor potency in structure-guided drug design. Multiple computational approaches exist to predict which waters should be targeted for displacement to achieve the largest gain in potency. However, the relative success of different methods remains underexplored. Here, we present a comparison of the ability of five water prediction programs (3D-RISM, SZMAP, WaterFLAP, WaterRank, and WaterMap) to predict crystallographic water locations, calculate their binding free energies, and to relate differences

in these energies to observed changes in potency. The structural cohort included an HIV protease structure, nine Bruton's Tyrosine Kinase (BTK) structures, and nine bromodomain structures. Each program accurately predicted the locations of most crystallographic water molecules. However, the predicted binding free energies correlated poorly with the observed changes in inhibitor potency when solvent atoms were displaced by chemical changes in closely related compounds.

Introduction.

Recently, targeting crystallographically observed water molecules for displacement or specific interaction to improve ligand affinity has proved successful in drug discovery. An array of computational methods now exist that attempt to predict which specific water(s) in a binding site should be targeted for displacement to achieve the largest gain in potency. Existing programs analyze water locations and energetics using different approaches ranging from simply scoring observed crystallographically observed water molecules, to grid-based sampling methods, to extensive use of molecular dynamics simulations to generate water positions and their corresponding energies (Table 1). Our interest in assessing water analysis tools was based on achieving a better understanding of hydration in active sites so that, ultimately, our ability to design molecules with increased potency, selectivity, binding kinetics, etc. can be achieved prospectively with proper use of these computational tools.

Table 1. Water placement tools used in this study.

	WaterMap[a]	SZMAP[b]	WaterFLAP[c]	3D-RISM[d]	Proasis WaterRank[e]
Method	MD simulation	Grid-based	Grid-based	Grid-based	Geometry-based
Scoring of X-ray water	N[f]	Y	Y	N	Y
Water prediction	Y	Y	Y	Y	N
Water score (unit)	ΔG	ΔΔG	ΔG_{wat} (kcal)	ΔG_{hyd} (kcal/mol)	Geometric score
Uses original PDB file	N (minimized)	Y	Y	Y	Y

[a] Maestro Version 2015-04. [b] SZMAP Version 1.2.0.7. [c] FLAP Version 2.2.0. [d] MOE Version 2015.10. [e] Proasis Version 3.
[f] WaterMap accepts X-ray water coordinates as starting points for MD simulations.

Molecular dynamics (MD) simulations are utilized by the program WaterMap.[1] In that case, a 2 ns simulation is run on a semi-constrained protein (only hydroxyl rotors allowed to move) within a box of waters to determine the preferred locations of water. Simulations can include static ligands with only hydroxyl rotors being free to move. Inhomogeneous solvation theory[2, 3] is then applied to determine the enthalpies and entropies of predicted waters. Recently, Carlson *et al.*[4] reported success with an unconstrained mixed-solvent molecular dynamics method that identifies water sites that are likely to be undisplaceable. Other simulation-based methods include Grand Canonical Monte Carlo (GCMC)[5], Just Add Water moleculeS (JAWS)[6] and double decoupling-Monte Carlo[7]. Monte Carlo methods may permit better sampling than MD-based methods but were not assessed in this study.

Grid-based methods include WaterFLAP[8], SZMAP[9], 3D-RISM[10, 11] and WaterScore (GIST)[12]. WaterFLAP (Fingerprints for Ligands And Proteins) locates water hydration sites using their proprietary GRID molecular interaction fields followed by scoring for hydrophobic and entropic character using their CRY and ENTR fields.[8] The implicit-water nature of using GRID technology is taken one step further with SZMAP, in which explicit waters are placed at grid points within the active site, followed by calculation of the energetics of each probe water relative to the same probe with charges removed, and separately, with van der Waals terms removed. The latter is termed a 'vacuum' probe. 3D-RISM (3D Reference Interaction Site Model) produces an approximate average solvent distribution around a rigid solute.[10, 11] Based on the density functional theory of liquids in the grand canonical ensemble, 3D-RISM produces an approximate average solvent distribution around a rigid solute using statistical mechanical methods. The evaluation of self-consistent equations generates solvent density maps that suggest the location of solvent molecules (waters). The contribution to the total solvation free energy can then be computed at each point on a grid allowing one to infer the displaceability of individual waters.

WaterRank[13, 14] provides a more enthalpic determination of existing waters in crystal structures but does not predict the location of such waters prospectively. WaterRank provides an assessment of water displaceability by statistically analyzing geometries of waters within existing crystal structures. Others tools such as AQUARIUS[15, 16], SuperStar[17, 18], and Consolv[19] are beyond the scope of this study.

Correlations between observed changes in potency and energetic calculations of displaced waters using a specific program have been reported. Abel *et al.* showed a respectable correlation (R^2=0.81) between calculations from an early version of WaterMap on 31 congeneric ligand pairs of inhibitors for Factor Xa spanning a range over 6 kcal/mol of relative free energies.[1] Beuming and colleagues used WaterMap on PDZ domains and found a correlation between the calculated free energies of explicit water molecules and the peptide ligands that displaced them.[20] Chrencik and colleagues used WaterMap retrospectively to explain the large potency increase with the installation of a nitrile-containing moiety on their pan-JAK inhibitor by identifying a high energy water in the region occupied by the nitrile group.[21] Laha and co-workers examined a thiazole-based series of CDK5 inhibitors, using WaterMap's assessment of two specific hydration sites to explain increased potency for ligands that displaced a high-energy water and decreased potency for ligands that displaced a relatively stable water[22]. In addition to these examples there are cases where WaterMap was used to obtain enrichment with virtual screening.[23, 24].

Comparisons of specific pairs of water calculations programs have recently appeared. Nguyen *et al.*, compared GIST to WaterMap using a series of factor Xa inhibitors.[25] Bodnarchuk *et al.* used N9-Neuraminidase to look at JAWS, GCMC, and double-decoupling.[26] Mason *et al.* applied a combination of GRID, WaterMap and SZMAP to a set of G-protein coupled receptors to assess the druggability of GPCR binding sites.[27] Bortolato and Mason analyzed the properties of the active sites and associated waters within protein-ligand complexes of 12 congeneric adenosine A_{2A} receptor antagonists as using a combination of WaterMap, SZMAP, GRID, and GCMC, concluding that the methods were complementary.[28] A comprehensive review on the role of water in the protein matrix recently appeared, in which water calculation tools are described in detail.[29] However, to our knowledge, no direct comparison of the ability of multiple programs to predict water locations and energetics, and map those predictions to differences in potency across multiple protein and ligand families have been reported.

The most challenging aspect of analyzing solvent positions in protein structures is predicting their energetic contribution to the complex. In ligand design, success of such predictions cannot be assessed on their own, and instead the quality of the prediction is indirectly read out as the relative affinity of a new ligand bearing atom(s) to displace a given water molecule compared to the reference ligand. This approach conflates prediction of water energy estimates with the quality of the ligand design that displaces it and therefore makes it difficult to objectively interpret. To address this inherent difficulty, we compared solvent energy calculations across 19 high resolution crystallographic protein-ligand complexes, nine of which were newly refined and deposited to the pdb. The cohort of crystal structures ranged from 1.0Å -2.1Å resolution, the majority of which were better than 2Å resolution. These structures were selected because ligands displaced specific solvent molecule(s) through conservative changes to the chemical structure of the ligand and with minimal shifts of binding pose relative to the protein. To assess the generality of our conclusions, we analyzed complexes from three protein families, Bromodomains (BRD), Bruton's Tyrosine Kinases (BTK), and an HIV protease.

In this study, the following software tools for predicting and scoring water molecules were evaluated: WaterMap (Schrödinger, Maestro Version 2015-04)[1], SZMAP (OpenEye, Version 1.2.0.7)[9], Proasis WaterRank (Desert Scientific, Proasis Version 3)[14], WaterFLAP (Molecular Discovery, FLAP Version 2.2.0)[8], and 3D-RISM (Chemical Computing Group, MOE Version 2015.10)[10, 11]. Three aspects concerning the accuracy of the programs were evaluated: (1) precision in the prediction of the location of crystallographically observed water molecules; (2) re-creation of the crystallographic water network using predicted water oxygen positions; and (3) correlation between predicted water energies and the observed structure-activity relationships (SAR).

Data Sets.

Crystal structures of protein-ligand complexes of HIV protease, BTK, and BRDs were selected to facilitate comparison of solvent networks by considering well-refined structures of better than 2.1Å resolution containing ligands with only minor chemical changes. In total, ten novel structures were determined to fill in the cohort. All of the structures crystals were grown by established methods and structures determined by molecular replacement. Accession codes and data and refinement statistics are provided in Supporting Information (Tables S1 and S2). Binding affinities, as measured in *in vitro* displacement assays, were available for all small molecule inhibitors, allowing a comparison of the energetics of water displacement to the effect on potency to be made.

Bromodomains (BRD). Four compounds (Figure 1a) co-crystallized with three proteins, BRD-9, BRD-4, and TAF-1, were selected for the analysis of displaced water molecules.[30] In these bromodomain structures, four left-handed alpha helices are packed in an antiparallel bundle. Two loop regions are present between helices αA and αZ (ZA loop) and αB and αC

(BC loop) and the acetylated lysines bind in the hydrophobic pocket created by these loop regions with their amides usually forming a direct hydrogen bond to a conserved asparagine located at the beginning of the BC loop (Figure 1b). The ligands in the bromodomain structures we studied possess a pyrrolopyridone core that also contains hydrogen bond interactions to this Asn100 (BRD9 numbering) through the carbonyl oxygen of the pyridone and the NH of the pyrrolo moiety. The tail extending from the pyridone portion of the core makes numerous vdW interactions with a neighboring lipophilic shelf in the protein that is formed primarily by Gly43, Phe44, and Phe45 and other nearby lipophilic residues. A total of nine Bromodomain structures were analyzed (Table 2, Figure 1c). Although the sequences are not identical, the topologies of the acetyllysine binding sites are identical with only minor motion of the protein heavy atoms observed. The binding sites were therefore superimposed. This afforded an ideal opportunity to assess the role of water displacement on ligand binding potency. Each compound displaced different water molecules within each protein structure, causing complete displacements of water molecules or rearrangements of the water network in most cases.

Details of the competition TR-FRET based binding assay used to assess **1**, **2**, **4**, and **5** have previously been described.[30, 31]

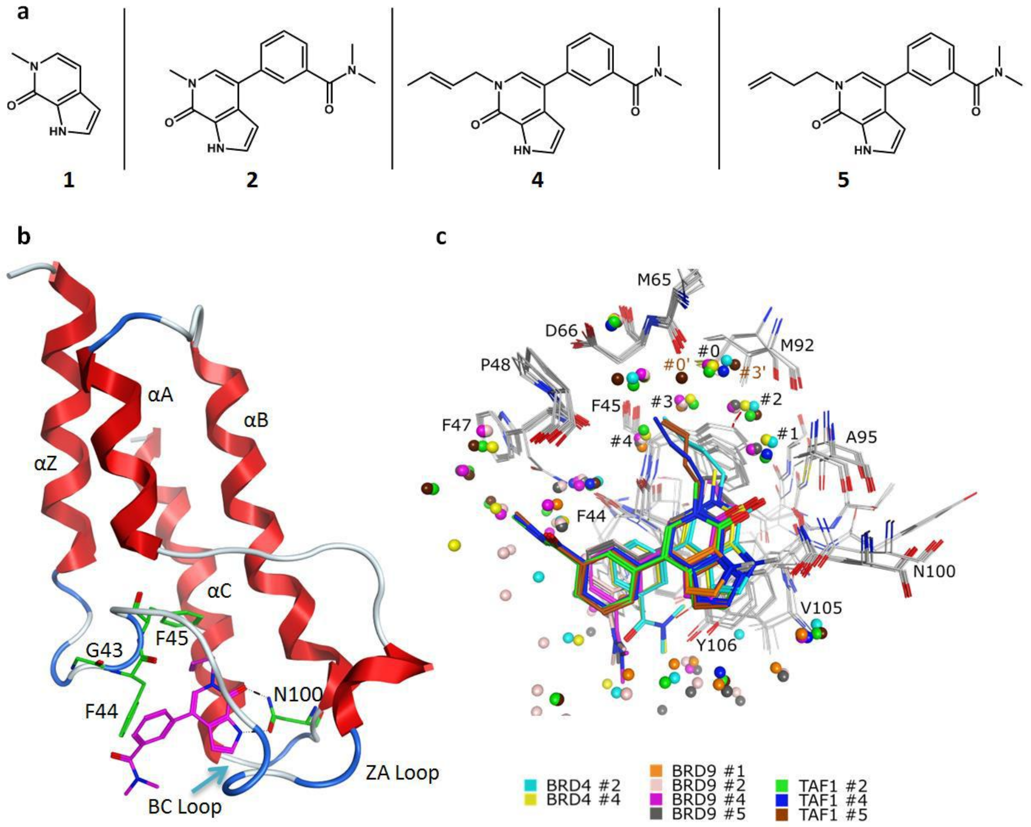


Figure 1. a) Co-crystallized BRD inhibitors. Compound numbering was taken from Crawford et al.[30] b) Crystal structure of the complex of compound **5** with BRD9. Conserved amino acids and adjacent secondary structural elements are labeled. c) Superposition of all 9

inhibitors showing consistency of binding mode, protein conformation, and water occupation. Waters chosen for the analysis are numbered. Amino acid numbering is based on BRD9.

Table 2. Bromodomain structures and their corresponding crystallographic water oxygen numbers, taken from Crawford et al.[30]. Binding affinities were also taken from Crawford et al.[30].

Compound	IC_{50} (μM)^a	Protein	X-ray Resolution (Å)	Observed Waters^b					PDB accession code
				#0	#1	#2	#3	#4	
2	0.092	BRD4(1)	1.45	x	x	x	x	x	5i80
4	0.46	BRD4(1)	1.40	c	x	x	d	-	5i88
1	13.7	BRD9	1.04	x	x	x	x	x	5i40
2	0.23	BRD9	1.15	x	x	x	x	x	5i7x
4	0.16	BRD9	1.45	x	x	x	x	x	5i7y
5	1.4	BRD9	1.03	x	x	x	x	x	6bqa
2	0.059	TAF1(2)	1.21	x	x	x	x	x	5i29
4	0.41	TAF1(2)	2.14	e	x	x	-	-	6bqd
5	0.046	TAF1(2)	1.49	f	x	x	g	-	5i1q

a) IC_{50} values were not significantly shifted by assay components, as the concentrations of the biotinylated competitor and bromodomain protein were well below the IC_{50} values.
b) Crystallographic waters present in the complexes are denoted by "x". Cells with dashes correspond to waters that were displaced by the ligand in the complex
c-g) Denotes waters with ≥ 1.2 Å shifts in position relative to water molecules in the reference X-ray complexes (compound 2 with BRD4(1) and TAF1(2)). Specific distances (in Å) were: c, 1.2; d, 2.9; e, 1.2; f, 1.3; g, 3.0.

Bruton's Tyrosine Kinases (BTK).[32–35] Nine BTK inhibitors (eight previously unpublished) were selected to examine their effect on the water network within the kinase ATP binding site (Figure 2a). All inhibitors bind in the same canonical orientation (Figure 2c), with the core heterocycle and linker NH forming two hydrogen bonds to Met477 of the hinge. An internal water network adjacent to this heterocycle, stabilized by interactions with the inhibitors, the gatekeeper Thr474, and Lys430 is observed across all complexes. The left-hand portion as shown in Figure 2b extends into solvent. As previously reported[36], these inhibitors induce a conformational rearrangement of the activation loop, creating a selectivity pocket formed in part by Gln412, Phe413, Asn526, and Tyr551. Germane to this analysis, the core heterocycles and protein active sites are well superimposable among the nine inhibitors with little protein motion observed between complexes. Eight water molecules were identified that are adjacent to the inhibitors and conserved across many of the X-ray complexes. Two of the eight water molecules are deeply buried in the binding site, whereas the remaining six form two clusters that interact with each other. Compound **7** contains a hydroxymethyl extension of the methyl in **6**, prepared to examine the effect on potency conferred by interactions with waters 3, 5, and 6 (Figure 2c), along with Asp539, and Lys430. The marginal 2x difference in potency suggests there is little to be gained by making these interactions. The **8-9** pair of compounds probed the extension of the right-hand side further into the selectivity pocket, potentially to displace water 8 in this region. The loss in potency of **8** relative to **9** suggests the potency difference is controlled partially by changes in the left-hand portion, coupled with a displacement of water 1 (for N-Me compound **8**), which is retained in the NH analog **9**. Analog **10** contains a hydroxyethyl as opposed to the hydroxymethyl in **11**, designed to probe deeper into the internal water cavity formed by waters 2, 5, and 6, potentially displacing one or more of them. Little difference in potency is observed between them, again suggesting there is little to be gained by disturbing the water network in this area. Compounds **12** and **13** (as well as **8-9**) contain a heterocycle NH versus N-Me to examine the effect of displacing water 1, which forms a water-mediated interaction in the NH heterocycle analogs (**9** and **13**) between inhibitor NH and the backbone carbonyl of Glu475. Both N-Me analogs (**8** and **12**) show decreased potency relative to their NH counterparts **9** and **13**, respectively, suggesting water 1 is relatively thermodynamically stable and difficult to displace. Finally, compound **14** contains a tricyclic right-hand heterocycle, designed to remove a hydrogen bond donor, along with an extension on the left-hand side to pick up additional interactions with the protein close to the solvent front. This latter compound is in early clinical development[37].

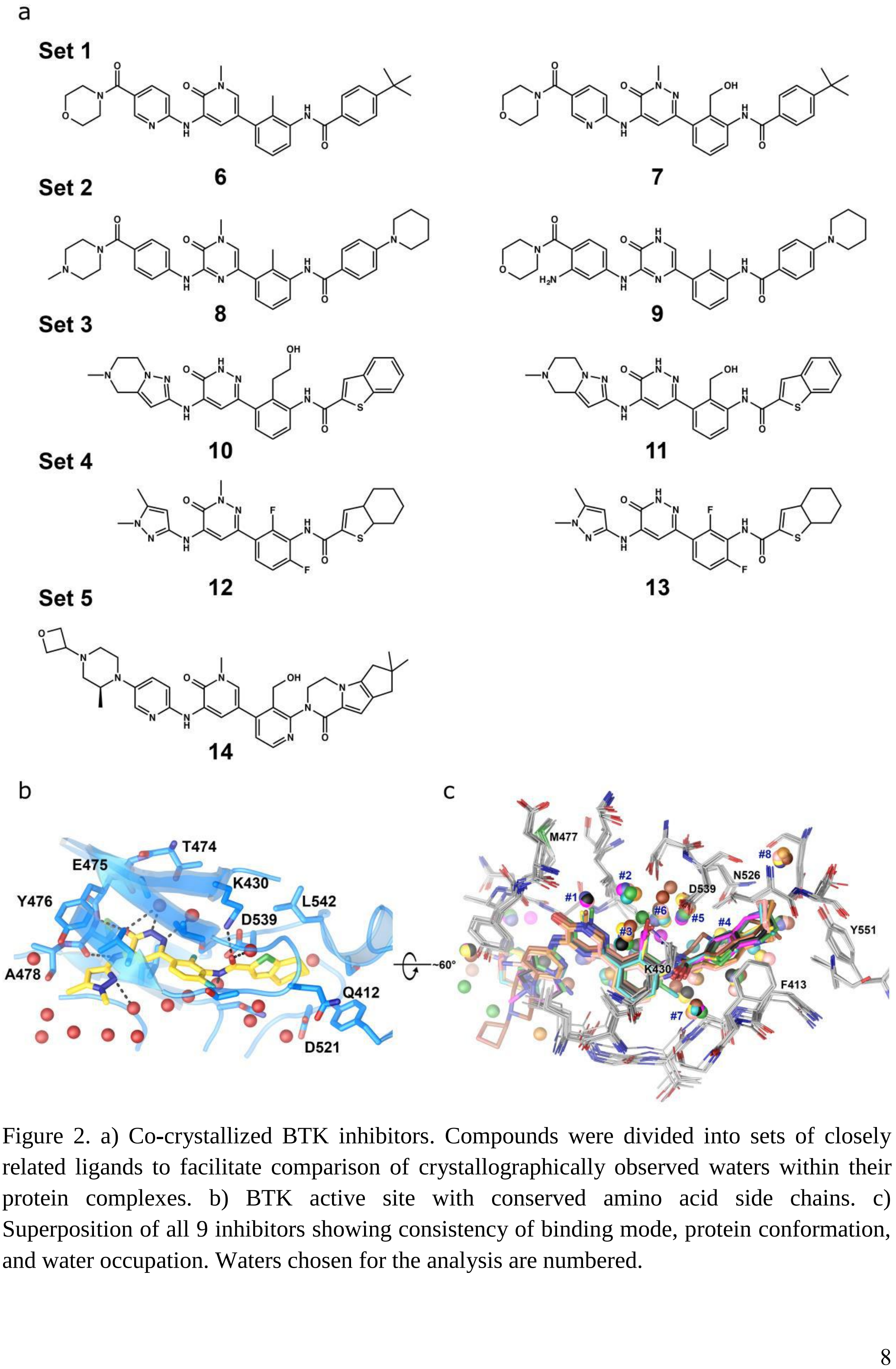


Figure 2. a) Co-crystallized BTK inhibitors. Compounds were divided into sets of closely related ligands to facilitate comparison of crystallographically observed waters within their protein complexes. b) BTK active site with conserved amino acid side chains. c) Superposition of all 9 inhibitors showing consistency of binding mode, protein conformation, and water occupation. Waters chosen for the analysis are numbered.

Table 3. BTK inhibitors, their corresponding water molecule IDs, and X-ray resolutions[a].

Compound	IC_{50} (μM)[b]	X-ray Resolution (Å)	Observed Waters[c]								PDB Accession Code
			#1	#2	#3	#4	#5	#6	#7	#8	
6	0.012	1.65	-	x	x	e	x	x	x	x	6aub
7	0.0064	2.05	-	x	x	f	x	-	x	x	6bik
8	0.145	2.01	-	x	x	g	x	x	x	x	6ep9
9[d]	0.0013	1.66	x	x	x	h	x	x	x	x	6aua
10	0.0045	2.15	x	x	x	x	x	x	x	x	6bke
11	0.0013	1.85	x	x	x	i	x	x	x	x	6bkh
12	0.005	1.70	-	x	x	x	x	x	x	x	6bkw
13	0.0016	1.40	x	x	x	x	x	x	x	x	6bln
14	0.00091	1.59	-	x	j	x	k	x	x	x	5vfi

a) Compound 13 was selected as the reference structure because all eight waters were present in the X-ray complex.
b) For assay details, see the S1 in the Supporting Information.
c) Crystallographic waters present in the complexes are denoted by “x”. Cells with dashes correspond to waters that were displaced by the ligand in the complex.
d) The ligand in this X-ray complex was modeled with two ring conformations. Conformation A was chosen for the present work. The water molecules of interest were identical between the two conformations.
e-k) These waters were ≥ 1.5 Å from the corresponding water molecules in the reference structure (compound 13). Specific distances (in Å) were: d, 1.71; e, 1.64; f, 1.69; g, 1.94; h, 2.10; i, 1.55; j, 1.71.

Human Immunodeficiency Virus (HIV) Protease. An HIV protease/inhibitor X-ray complex (1kzk[38]) was selected to analyze the effect of a single water molecule that is well integrated in the binding site (Figure 3). This water (HOH-A-1037) is nearly ideally coordinated, forming two interactions each to the protein and the bound ligand. Since it has been proven difficult to replace this water molecule, with no observed gain in binding affinity in doing so[38, 39], this water molecule should be rated as thermodynamically stable by all water prediction programs.

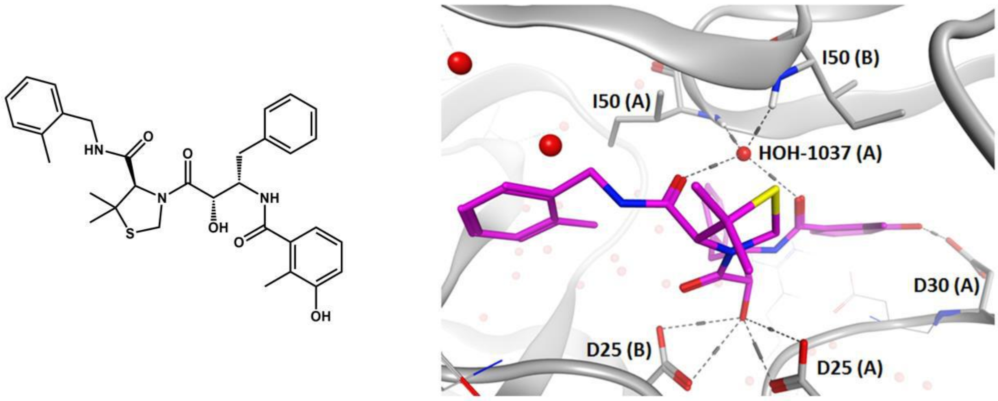


Figure 3. HIV protease inhibitor and corresponding protein-ligand interface with interacting water molecules included (PDB accession code 1kzk). The key water analyzed in this study is labeled, along with other key residues that form hydrogen bonds to the ligand. Numbers in parentheses denote the protein chain.

Methods.

Selection of Crystallographic Water Molecules. In a previous report[30], five specific water molecules that mediate protein-ligand interactions in BRD structures were analyzed in relation to the observed structure activity relationships. These five waters were selected for the present study. Water molecules in BTK structures were selected based on their proximity to the bound ligand, their invariance across crystal structures, and their displacement by ligand modifications across selected matched pairs of inhibitors. Specifically, two water clusters (Figure 2c, waters numbered 1,2,3 and 4,5,6) that were affected by ligand alterations were selected. Additionally, two water molecules that appear in all complexes in confined areas of the protein adjacent to the inhibitors (#7 and #8) were chosen due to their presumed thermodynamic stability. As stated previously, for HIV protease[38], a single well integrated water molecule that mediates interactions between ligand and flaps on the protein was chosen. Because the crystal structures emanated from multiple crystallographers across multiple labs, all water molecules were checked for electron density manually as well as with an automatic criterion, called EDIA (Electron Density of Individual Atoms).[40, 41],[42]

Evaluation Criteria. Four different aspects concerning computational placement of water molecules were analyzed to assess the accuracy of each program: (1) Distance between predicted and crystallographically observed water oxygens; (2) Number of predicted and crystallographically observed water molecules in an ‘area of interest’; (3) Difference between the water distance networks formed by the predicted and crystallographically observed water oxygens, and (4) Correlation of the predicted water energies with the experimentally observed SAR.

Distance of Predicted to Crystallographically Observed Water Oxygens. Each predicted water oxygen was assigned to the closest available crystallographically observed water oxygen. Thus, every crystallographic water oxygen in this area had only one predicted water oxygen assigned to it. After each crystallographic water oxygen had a predicted water oxygen matched to it, the distance between the predicted and crystallographic water oxygen was measured.

Number of Predicted Water Molecules. The overall number of predicted water molecules in an ‘area of interest’ was counted. The area of interest was defined for each protein target to enclose the protein-ligand interfacial region and associated crystallographically observed key waters that were selected for analysis (Figure 4). Predicted waters that fell outside this area of interest were classed as ‘missed’ by the programs. The radii of the spheres defining these areas were set to accommodate crystallographic waters plus a 2.5 Å tolerance.

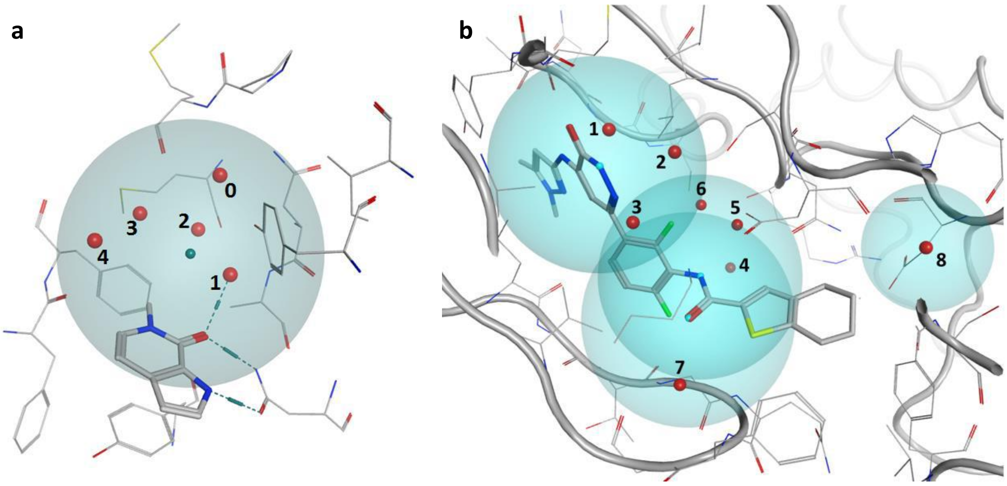


Figure 4. a) An ‘area of interest’ used to define predicted water molecules in BRD and TAF structures, illustrated using the BRD9/**1** complex. A cyan point is shown that depicts the sphere center. Sphere radius was chosen to be 5 Å to encompass conserved internal waters (numbered 0-4) observed across all crystal structures. b) ‘Areas of interest’ for placed water molecules in BTK, illustrated using compound **13**. These areas were defined by placing three 4.5 Å radius spheres around points placed on the amide nitrogen and oxygen atoms, and the pyridazinone ring nitrogen (atoms highlighted with cyan points). Sphere radii were chosen to encompass conserved internal waters. Water molecule 8 (far right edge) was included using a separate 2.5 Å radius sphere due to its distance from the remaining waters.

Assignment of equivalent water identities became increasingly difficult in cases where the program predicted large numbers of waters within the area of interest. In some instances, some predicted waters could not be assigned and were excluded from tabulations.

Re-creation of the Water network. Measuring distances between individual water oxygens, and counting the number of predicted versus crystallographic water molecules are not enough to judge the quality of a water prediction program. Therefore, the observed and predicted water-water distance networks were analyzed by measuring pair-wise distances between the oxygen atoms of relevant water molecules. These distance networks were then compared to each other using a root-mean squared-deviation (RMSD) metric. Specifically, the distances between the crystallographic water oxygens were compared to the distances between the predicted water oxygens within each network. Every crystallographic water was assigned to one predicted water. In cases where the program failed to predict a nearby water oxygen position (nearest predicted water was >2.5 Å from a crystallographic position), the next closest unassigned water oxygen was used, resulting in an increased RMSD in these cases. Thus, smaller RMSD values reflect greater fidelity in the recreation of the water network.

Energetic Contribution of Water Molecules and SAR Consistency. Energies were calculated for predicted and crystallographic water molecules. Calculating energetics of observed crystallographic waters allowed a direct assessment of the differences between the programs to be made. Energetics for crystallographic coordinated waters could be retrieved from

SZMAP, WaterFLAP, and WaterRank. We analyzed the relationships between the calculated energies and the experimentally observed SAR, and examined the intercorrelation of the energies between the programs. Evaluations were run on the protein-ligand complex by removing crystallographically observed waters, retaining the ligand, and allowing the programs to place calculated waters back into the protein-ligand complex. Evaluations were also run using the holo structure, where crystallographically observed waters *and ligands* were removed, leaving an empty active site for the programs to place calculated waters back into. WaterRank could not be evaluated by these approaches because it does not offer water placement.

Program Options. All software (Table 1) was used with default options. An aim of this study was to compare results from the programs as supplied 'out of the box' without additional adjustment of internal program settings. The only exception was the application of WaterMap to BTK complexes. To reduce the computational time to a manageable level, the active site around the larger BTK inhibitors was reduced in size from the default 10.0 Å to 8.0 Å surrounding the inhibitor, and the simulation time was decreased from 2 ns to 1.5 ns. All complexes were prepared using the procedures recommended for each program before the water placement and/or energy predictions were run.

Results.

Program Output. Each program generates a different output. Therefore, the numbers have to be interpreted appropriately. All programs, except WaterRank, generate water energies. Negative energy values indicate waters in a thermodynamically stable environment whereas positive energy values indicate an unfavorable and therefore rather unstable water environment. Therefore, waters with negative energies are thought to be more difficult to displace than those with positive energies. It should be noted that absolute numbers that resulted from the different programs cannot be compared directly. Some estimate ΔG values (WaterMap), some ΔΔG (SZMAP), and others ΔG_{wat} or ΔG_{hyd} (WaterFLAP and 3D-RISM). Therefore, only relative numbers are compared throughout this study.

WaterRank produces a geometric score for each observed crystallographic water where the hydrogen bond distances and angles between the hydrogen bond partners are compared to ideal values of 2.8 Å and 109.5° (tetrahedral angle).[14] Only protein atoms are considered as hydrogen bond partners while ligand atoms are excluded. A maximum of two donors and two acceptors are allowed as potential water molecule partners. Higher WaterRank scores reflect smaller deviations from ideal coordination geometry. An ideal, tetrahedrally coordinated water molecule is given a maximum score of 6.0. WaterRank scores are classified into categories according the likelihood of the waters to be displaced: 'easy to displace' (scores 0 – 2.3) and 'possible to replace' (scores 2.3 – 4.0).[14, 43]

Bromodomains (BRD). Five water molecules were previously identified as conserved across all Bromodomain structures (Figure 5).[30] These water molecules were not only conserved in ligand-bound structures, but also in the apo structures of the proteins. They were therefore analyzed for their placement by each program as well as their energetic contributions to binding affinity.

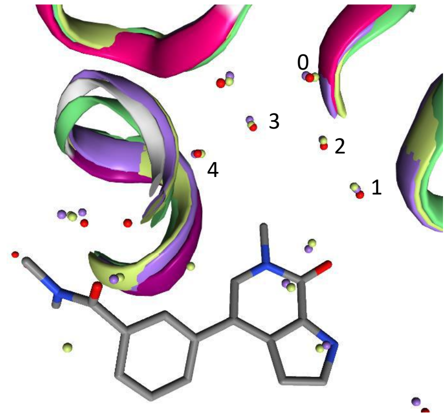


Figure 5. Overlay of TAF1(2) structures using ASCONA[44, 45]. The TAF1(2)/compound **2** complex is shown (magenta ribbon, crystallographically observed water oxygens as red spheres) superimposed onto the apo structures from the following complexes taken from the PDB: 3uv5 (yellow), 3aad (green), and 1eqf (purple). Water oxygens are numbered according to reference 30; note their near invariant positions across the complexes.

Distance to Crystallographic Water Molecules. The distance to crystallographic water molecules provides information about the accuracy of the predicted water locations by each method (Figure 6 and S5). WaterMap achieved greater than 70% accuracy at correctly placing water molecules within 0.5Å of their observed positions (~90% placed within 1Å, Figure 6). However, water molecule #1 (Figure 7), which is buried inside the pocket and mediates interactions between protein and ligand, was not placed in four out of the nine simulations (distance from the crystallographic position to the nearest predicted water molecule was >2.5Å). No underlying cause for this failure was identified. A similar environment surrounding that water position was present in structures where the water was accurately placed (Figure S6 and Table S3). WaterFLAP, SZMAP, and 3D-RISM placed water molecules within 1Å of their crystallographically observed positions with a success rate of approximately 80%, 70%, and 60%, respectively.

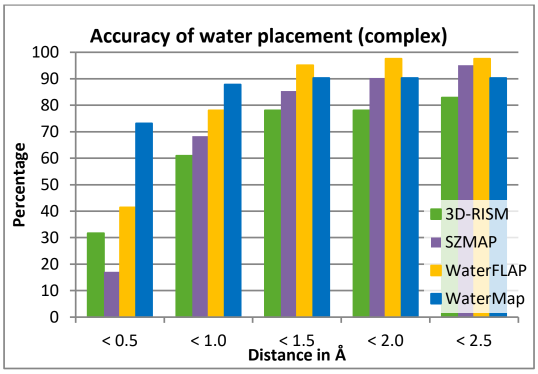


Figure 6. Distance of predicted water molecules to the corresponding 41 crystallographic water molecules across all nine Bromodomain complexes.

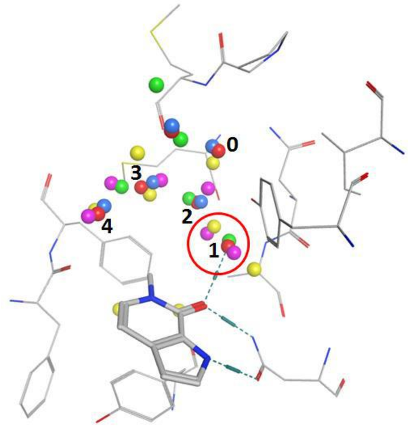


Figure 7. Predicted versus observed water molecules for the Bromodomain BRD9/**1** complex. Crystallographic water oxygens are in red. Hydrogen bonds between **1**, protein, and a crystallographic water are shown as dashed lines with cylinders. Water oxygen atoms placed by WaterFLAP (yellow), SZMAP (pink), 3D-RISM (green), and WaterMap (blue) are superimposed. WaterMap did not predict the tightly bound water molecule #1 (circled).

Number of predicted water molecules. An ‘area of interest’ was defined using a sphere that included internal crystallographically observed water molecules that were conserved across all Bromodomain complexes. Specifically, a carbonyl oxygen common to all BRD ligands and a conserved amide backbone oxygen (Met65 sidechain in the case of BRD9) were chosen to define the center of the sphere (Figure 4). Inclusion of surface exposed water molecules would give rise to too much variability in water placement due to too few contacts these waters make with the protein.

WaterMap predicted a similar number of water molecules as were present in the crystal structures, with a total of 43 out of 48 crystallographic waters reproduced (Table 4). 3D-RISM generated a similar number of placed waters (42 out of 48), with too few water molecules predicted in some structures (TAF1 #4) and too many in others (BRD4 #2). WaterFLAP on the other hand predicted more water molecules in all structures. The highest number of water molecules was predicted by SZMAP, in some cases double the number of water molecules in the ‘area of interest’ than were observed in the crystal structures (Figure 8). SZMAP placed multiple waters in close proximity to one another near crystallographic observed water molecules. Clustering these groups of placed water molecules to arrive at a consensus placed water position might be a way to post process SZMAP results to arrive at more realistic predicted water positions. The SZMAP result illustrates the difference in the aim of the SZMAP program compared to other methods. SZMAP was not developed to place specific water molecules, but rather to indicate areas where a ligand might be modified/expanded to capture predicted water sites. The fact that SZMAP placed multiple waters in closely packed clusters complicated the interpretation of results from this program and made direct, clear comparisons with the other programs difficult.

Table 4. Number of crystallographically observed and predicted water molecules by each program within the 'area of interest' for BRD and TAF complexes (see Figure 4).

Compound	Protein	Crystal Structure	3D-RISM	SZMAP	WaterFlap	WaterMap
2	BRD4(1)	6	8	11	9	6
4	BRD4(1)	5	4	12	6	5
1	BRD9	6	5	6	7	5
2	BRD9	6	5	7	7	6
4	BRD9	6	5	5	9	5
5	BRD9	6	5	8	9	5
2	TAF1(2)	6	4	12	7	4
4	TAF1(2)	3	2	5	6	3
5	TAF1(2)	4	4	8	7	4
Sum		48	42	74	67	43

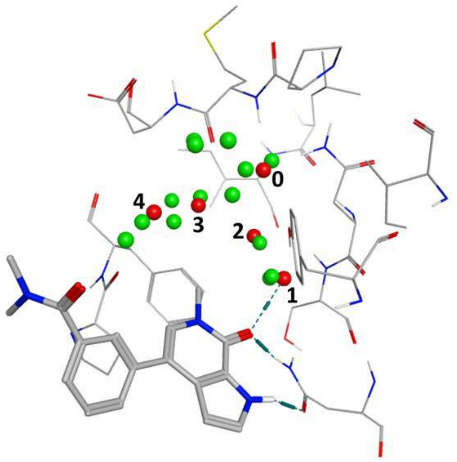


Figure 8. Crystallographically observed key water oxygens (in red, numbered) and SZMAP-placed water oxygens for the TAF1(2)/**2** complex. Clusters of closely placed waters are apparent.

Re-creation of the water distance network. Pairwise distance matrices between crystallographically observed water oxygens were created for all Bromodomain complexes as well as water oxygens placed by the different programs (Figure 9 and Figures S7-S9). For each protein structure, the matrix of placed water oxygens was compared to the corresponding matrix from the crystal structures. As an objective criterion for comparing the re-creation of water networks by each program, the RMSD was calculated for each pair-wise distance difference (Table 5, Table S4).

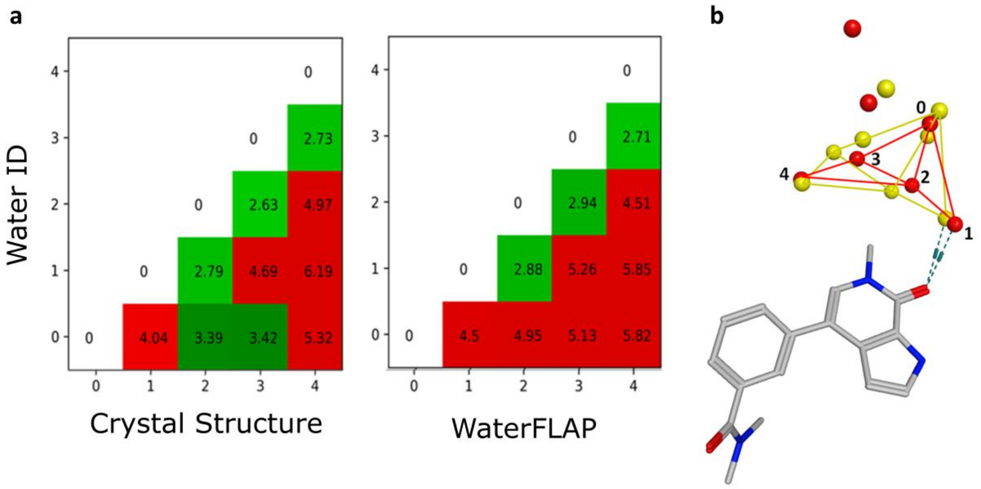


Figure 9. Re-creation of the water network. (a) Pairwise distance map of observed crystallographic (left) and WaterFLAP placed (right) water molecules present in the BRD4 complex with compound **2**. Numbers in the matrix are the pairwise distances in Å. (b) Overlay of the crystallographically observed (red) and WaterFLAP predicted (yellow) water oxygens and network from this complex.

WaterFLAP and WaterMap performed similarly well in recreating the observed water networks in the Bromodomain complexes. Each produced the smallest RMSD for four out of nine targets, with similar average RMSDs of 0.61 and 0.64 respectively. For complexes where WaterMap could not place water molecule #1 accurately, the RMSD values were relatively high, e.g., TAF1 with **2** (RMSD of 1.43). SZMAP achieved an average RMSD below 1.00 for seven out of nine targets and an overall RMSD of 0.84, with only two out of 41 placed water molecules greater than 2.5 Å from the crystallographic versions. Note that these results are directly connected with the number of placed water molecules. As greater numbers of waters were predicted by a program, the probability that one of them was proximal to a crystallographic water increased. At the same time, an inaccurately placed water molecule that was chosen to match the crystallographic water molecule could lead to a distorted, inaccurately created water network. With these caveats in mind, we found that 3D-RISM produced the highest variation from the crystallographically observed water networks. The lowest RMSD obtained from 3D-RISM calculations (BRD4 with **2**) was higher than the overall RMSD averages of the other tools. Interestingly, placing water molecules in the holo structure of the proteins, i.e. the protein without the ligand present, led to much better results for 3D-RISM (Table S4). Using the holo structures, the water networks could be re-created using 3D-RISM placed water molecules with an RMSD below 1.00 in all cases. The overall RMSD for 3D-RISM (0.79) for holo structures was superior to that of SZMAP (0.93) and WaterFLAP (0.86) calculations. WaterMap on the other hand achieved the same accuracy (overall RMSD of 0.64) in re-creating the water network with placed water molecules without the ligand present.

Table 5. RMSD of pair-wise distance matrices between crystallographic and computationally placed water molecules in the Bromodomain protein-ligand structures. Numbers in parentheses (x/y) denote the number of placed water molecules x >2.5 Å away from crystallographic water molecules y.

Protein	Compd	Crystal Structure	3D-RISM		SZMAP		WaterFLAP		WaterMap	
BRD4	2	0	0.87		0.44		0.81		0.19	
BRD4	4	0	1.38		0.61		0.40		0.07	
BRD9	1	0	1.16	(1/5)	0.88	(1/5)	1.58		0.97	(1/5)
BRD9	2	0	0.95	(1/5)	0.91		0.58		0.24	
BRD9	4	0	0.96	(1/5)	1.59	(1/5)	0.65	(1/5)	1.01	(1/5)
BRD9	5	0	0.93		1.23		0.66		1.02	(1/5)
TAF1	2	0	2.64	(1/5)	0.55		0.38		1.43	(1/5)
TAF1	4	0	3.37	(2/3)	0.39		0.13		0.69	
TAF1	5	0	1.54	(1/4)	0.95		0.33		0.18	
Averages			1.53	(7/41)	0.84	(2/41)	0.61	(1/41)	0.64	(4/41)

Energetic Contribution of Waters and Consistency with SAR. An overlay of all BRD9 structures (Figure 10) shows that the expansion of the ligand from **1** to **2** to **4** and **5** leaves the water network undisturbed. The hydrophobic tails of **4** and **5** extend into the back of the pocket, leading to a slight shift of Phe45, Tyr106, and Ile113 for some complexes. Therefore, the observed differences in binding affinities among these ligands cannot be 100% directly compared to the calculated water molecule energetics due to these slight shifts in protein conformation. However, since there is virtually no difference in the water network upon binding of the different ligands, the energies of the water molecules should remain fairly similar for all BRD9 structures. This assumption was analyzed using two different approaches: (1) scoring the observed crystallographic water molecules with SZMAP, WaterFLAP and Proasis WaterRank; and (2) scoring predicted water molecule positions with 3D-RISM, SZMAP, WaterFLAP, and WaterMap.

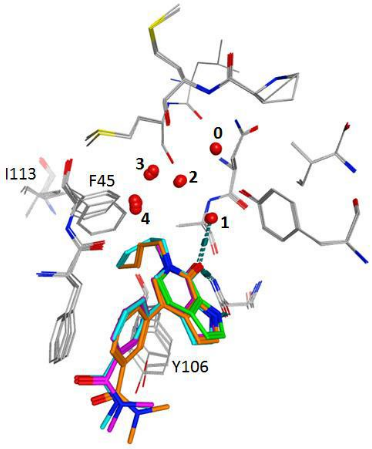


Figure 10. Overlay of BRD9 complexes showing an undisturbed network of conserved water oxygens (red spheres). Compound **1** is green, **2** is magenta, **4** is orange, and **5** is cyan. Residues whose conformations are somewhat altered are labeled.

Water molecules #0 to #3 all received the same WaterRank score for the different BRD9 structures (Figure 11; WaterRank graph). Water molecule #4 had two scores of 2.3 (BRD9 with compounds **1** and **5**) and two of 3.8 (BRD9 with **2** and **4**). These scores are consistent with the unchanged X-ray water network observed across all BRD9 structures. The energies for SZMAP-scored water molecules across BRD complexes were very similar with a somewhat higher variation noted for water molecule #4 (Figure 11; SZMAP graph). WaterFLAP-calculated energies showed greater variation among BRD complexes (Figure 11, see the WaterFLAP graph). However, the overall energy profile was similar. A comparison of the different tools revealed similar results between WaterRank and WaterFLAP. Both tools generated similar water molecule rankings, i.e., water molecule #0 received higher scores than #1. SZMAP on the other hand scored water molecules #0 and #1 similarly.

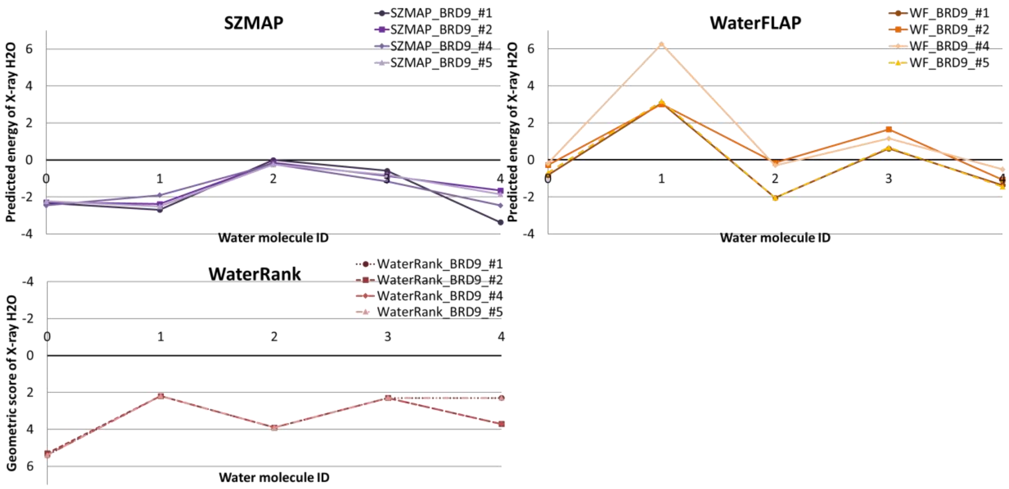


Figure 11. Predicted energies of crystallographic water positions. 3D-RISM and WaterMap do not allow the scoring of crystallographically observed water molecules, so they were not included.

The predicted energies for placed water molecules were less consistent and quite variable (Figure 12). The ranking of water molecules across BRD complexes according to their predicted energies even varied within the same program. The overall shapes of the graphs varied in many cases from complex to complex; this was true for each program tested. For example, 3D-RISM predicted water molecule #1 to be slightly unfavorable and water molecule #2 to be more unfavorable in BRD9 with compounds **1** and **5**, but with **2** the opposite result was seen. For compound **4**, both water molecules were scored nearly the same. It appears that in general, slight changes in the predicted water molecule positions seemed to profoundly affect the calculated energies. Moreover, only a limited trend between the different tools could be observed. 3D-RISM predicted water molecule #1 to be very favorably contributing to binding, whereas WaterMap and WaterFLAP rated it as unfavorable.

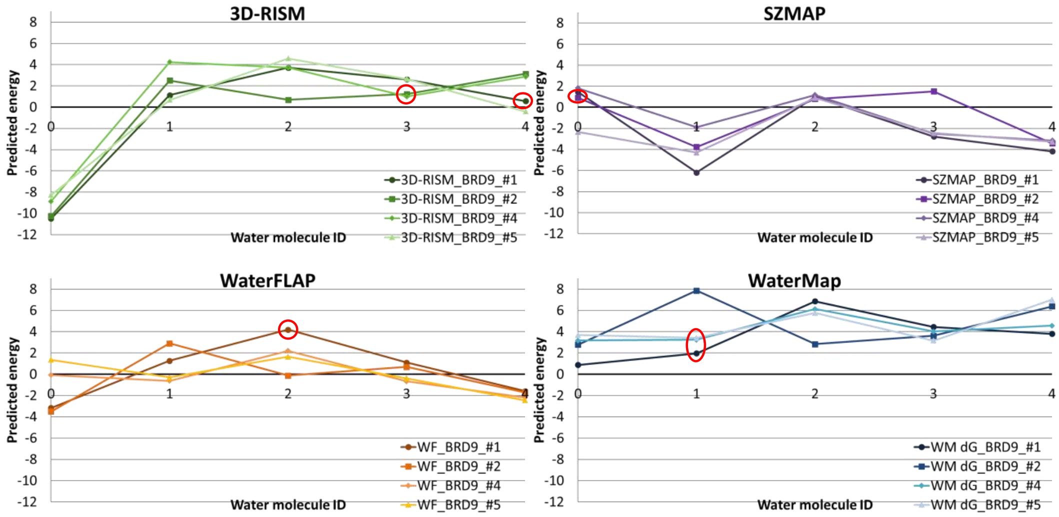
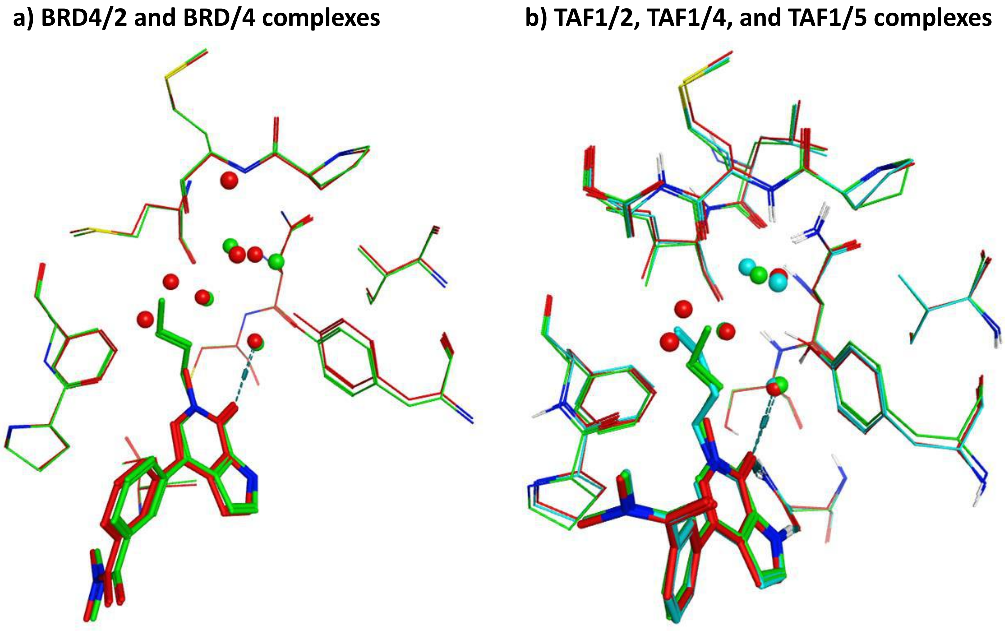


Figure 12. Predicted energies of placed water molecules. WaterRank can only score observed crystallographic water molecules so it is not included here. Waters placed >2.5 Å apart from the corresponding crystallographic water are circled in red. Connecting lines are for visualization purpose only.

Compared to the BRD9 structures, the ligands in BRD4 and TAF1 extend further into the water network, displacing water molecules and disrupting the water network (Figure 13).


a) BRD4/2 and BRD/4 complexes

b) TAF1/2, TAF1/4, and TAF1/5 complexes

Figure 13. Overlay of a) two BRD4 (left); and b) three TAF1 (right) crystal structure complexes. Compound **2** and associated crystallographic water oxygens are in red; **4** is green, and **5** is cyan.

The binding of compound **4** to BRD4 decreases compared to **2** (Figure 14). Upon binding of **4**, water molecule #4 gets displaced and water molecules #0 and #3 are shifted. Extending the ligand (compare **2** to **4**) also leads to a drop in affinity for TAF1. In this complex, water molecules #3 and #4 are displaced and #0 is shifted. A shift of the double bond in the pyridone N-side chain from the beta-gamma to the terminal position (see **4** and **5**) results in a shift rather than a displacement of water molecule #3, leading to increased affinity (Figure 14).

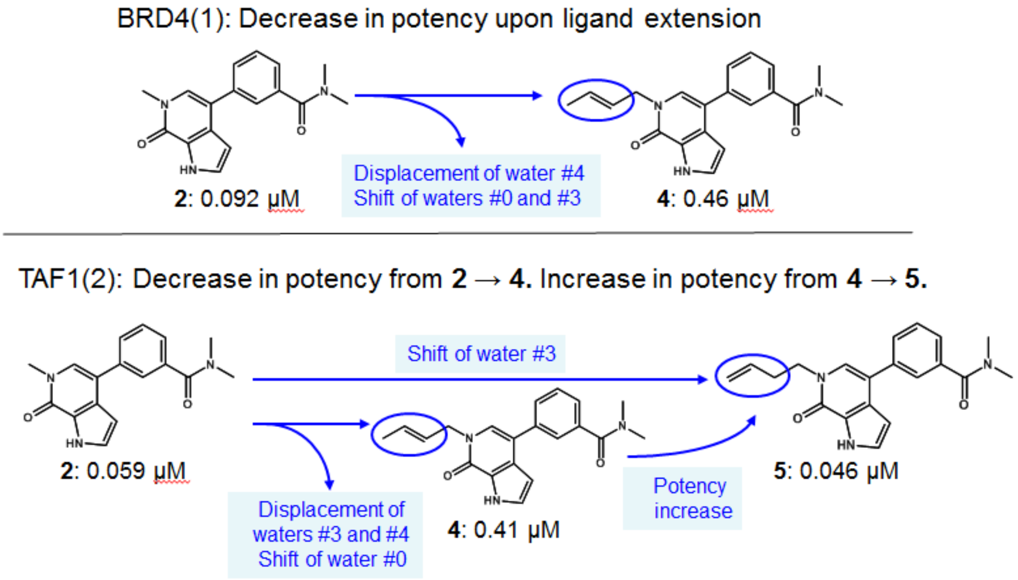

BRD4(1): Decrease in potency upon ligand extension
2: 0.092 μM
Displacement of water #4
Shift of waters #0 and #3
4: 0.46 μM
TAF1(2): Decrease in potency from 2 → 4. Increase in potency from 4 → 5.
Shift of water #3
2: 0.059 μM
Displacement of waters #3 and #4
Shift of water #0
4: 0.41 μM
Potency increase
5: 0.046 μM

Figure 14. Affinity and water molecule changes for BRD4 and TAF1 compounds.

The predicted energies for the water molecules were analyzed in two ways to draw conclusions related to the observed changes in affinity. The calculated energies for individual water molecules and the overall energy change for the entire water network were analyzed. Because some BRD ligands displaced waters on binding, comparisons between the observed changes in potency and the energies of the displaced waters could be made (Figure 14). Water molecule #3 was expected to be contributing favorably to the overall binding affinity, because its displacement in TAF1 with **4** led to a decrease in affinity. However, its shift (but not displacement) in TAF1 with **5** led to a potency increase. Because reduced affinity is seen when water molecule #4 is displaced while water molecules #0 and #3 are only shifted, water molecule #4 was also expected to contribute favorably to the binding affinity.

WaterRank scores for the crystallographic water molecules #3 and #4 in BRD4 and TAF1 are right at the edge of their classification ranges. Each of these four water molecules received a WaterRank score of 2.3 indicating they are between ‘easy to displace’ (scores 0 – 2.3) and ‘possible to replace’ (scores 2.3 – 4.0, Table 6). Both water molecules were scored favorably by SZMAP. SZMAP predicted water molecule #4 to be even more stable than water molecule #3. However, since water molecule #4 is displaced in all BRD4 and TAF1 structures, we would have expected water molecule #3 to be energetically more favorable than #4. The scores of crystallographic water molecules were more divergent for WaterFLAP, varying from unfavorable (BRD4 water molecule #3) to favorable (BRD4 water molecule #4).

Table 6. Predicted energies for crystallographic water molecules.

Protein	Compound	Water ID	WaterRank	SZMAP	WaterFLAP
BRD4	2	3	2.3	-0.87	0.82
BRD4	2	4	2.3	-1.33	-0.99
TAF1	2	3	2.3	-0.99	-0.18
TAF1	2	4	2.3	-1.83	-0.32

Similar to the RMSD evaluation of predicted water positions, the energy scores of the placed water molecules were highly diverse (Table 7). 3D-RISM predicted water molecules #3 and #4 to be unstable across all complexes, whereas SZMAP scored all waters to be favorably contributing. WaterFLAP rated water molecule #3 as unstable, whereas #4 had a favorable contribution. WaterMap scores were all unfavorable. Water molecule #4 was even more unfavorable than #3 across these complexes.

Overall, the individual water energies were not consistent with the corresponding SAR. Additionally, the results differed greatly for all programs, with no clear trend being discernable between the calculated energies and potency.

Table 7. Predicted free energies for placed water molecules in BRD and TAF1 complexes.

Protein	Compound	Water ID	3D-RISM	SZMAP	WaterFLAP	WaterMap (ΔH)
BRD4	2	3	2.13	-1.92	0.64	2.08 (-2.27)
BRD4	2	4	2.38	-1.73	-1.62	8.38 (-3.89)
TAF1	2	3	5.45	-1.97	0.35	3.96 (-0.20)
TAF1	2	4	0.28	-4.59	-1.96	7.05 (-2.96)

Examination of the water network in TAF1 shows that the change from **2** to **4** disrupts the water network **-** the extensive hydrogen bond network of the water molecules to the protein surface is broken (Figure 15). The change from **4** to **5** re-established this water network **-** water molecules now form hydrogen bonds to each other, reaching out to water molecules at the protein surface. The change of the double bond from the beta-gamma to the terminal position of the hydrophobic chain (**5** to **4**) alters the conformation of the ligand in a way that interrupts the hydrogen bond network in this rather narrow area of the pocket. To analyze this, we calculated the average score of the water network for each structure and recorded the change in the average water score for the different compounds (Table 8).

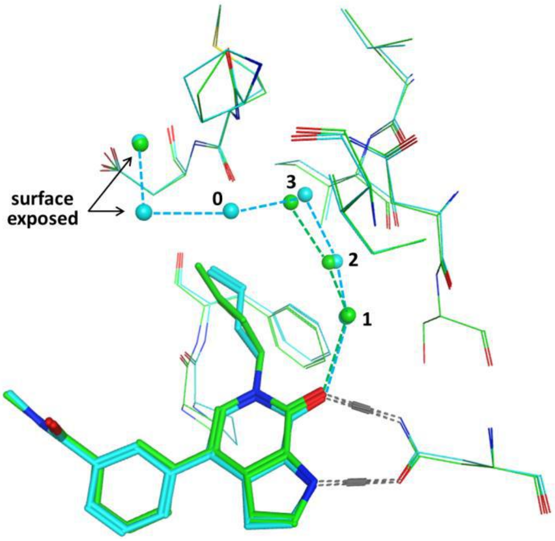


Figure 15. Overlay of TAF1 with compounds **4** (green) and **5** (cyan). Protein-ligand hydrogen bonds are shown as grey lines containing cylinders. A well-ordered water network extending from ligand to the protein surface is seen for **5** (cyan dashed lines), while **4** displaces water 0 and one of the surface interfacial waters, disrupting this network.

The average score of the water network in TAF1 should be favorable for **2**, decrease for **4**, and increase again for **5**. This trend could only be observed for the average water network score predicted by WaterMap (Table 8). However, the average WaterMap scores were extremely unfavorable, i.e. the whole water network was calculated to be unstable. SZMAP and WaterFLAP average water network scores showed the correct trend from **2** to **4** (a decrease in the energy contribution), but failed to identify the gain in energy on binding of **5**. Conversely, 3D-RISM showed the correct trend of a more favorable water network for **5** than **4**, but in general all predicted average energies were calculated to be unfavorably contributing to the binding affinity. In short, an average 'happiness' of the calculated water network was seemingly unrelated to the observed change in potency of the BRD and TAF ligands.

Table 8. Average water energies for TAF1 compounds. Green arrows denote an expected increase in energy based on potency difference that is correctly predicted by a given calculation. Red arrows show expected decrease in energy based on SAR that is correctly predicted by a given calculation. The absence of an arrow denotes a predicted change in energy that is inconsistent with the observed change in potency.

Protein	Compound	3D-RISM	SZMAP	WaterFLAP	WaterMap (ΔH)
TAF1(2)	2	2.47	-4.00	-1.18	5.20 (1.69)
			↓	↓	↓
TAF1(2)	4	1.98	-3.85	-1.00	7.16 (4.27)
		↓			↓
TAF1(2)	5	0.30	-3.06	-0.78	6.25 (1.54)

Bruton's Tyrosine Kinase (BTK). Eight conserved water molecules were analyzed for their correct placement and energetic contributions.

Distance to Crystallographic Water Molecules. WaterMap predicted water molecules most accurately in BTK structures, with more than 70% placed within 0.5 Å of the crystallographically observed waters and more than 90% within 1.0 Å (Figure 16 and Figure 17). All other methods achieved about 60% accuracy within 1.0 Å. For predictions on holo structures (Figure S10), the accuracy of WaterMap decreased more than for the other tools. WaterMap and WaterFLAP were nearly equal in their accuracies of predicting water positions in holo structures.

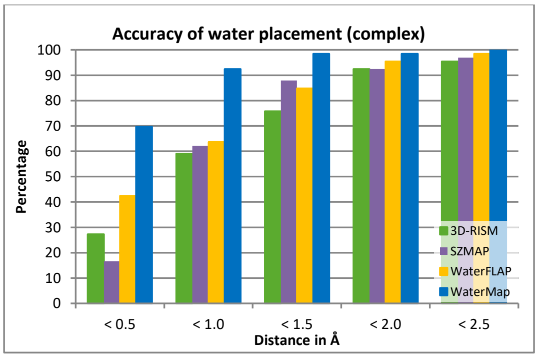


Figure 16. Distance of predicted water molecules to the corresponding 66 crystallographic water molecules within all nine BTK complexes.

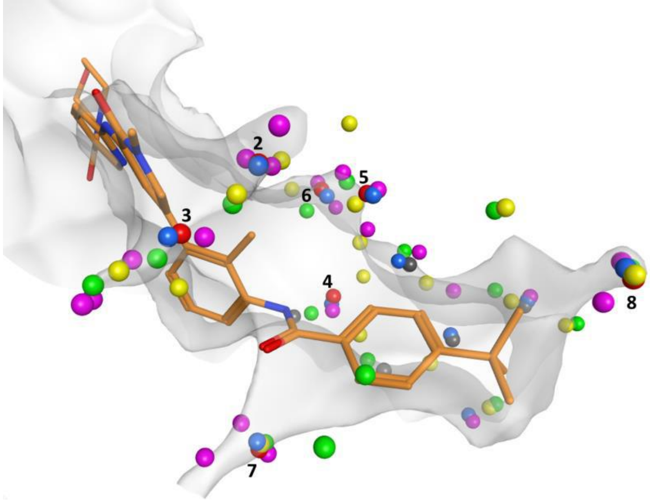


Figure 17. Predicted versus observed water molecules for BTK/**6** complex. Crystallographic water oxygens are shown in red and labeled by water number. Water oxygen atoms placed by WaterFLAP (yellow), SZMAP (purple), 3D-RISM (green), and WaterMap (blue) are superimposed. Additional crystallographic water oxygens that were not analyzed in this study are shown in grey. A solvent accessible surface has been added

Number of predicted water molecules. The ‘area of interest’ in BTK structures was defined using three spheres of 4.5 Å radius around conserved ligand atoms and an additional sphere of 2.5 Å radius around the isolated water #8 (Figure 4b).

The number of placed water molecules by WaterMap in the ‘area of interest’ showed the best agreement with the number of crystallographically observed water molecules (Table 9). The second most accurate was 3D-RISM with a total of 80 placed water molecules, while SZMAP and WaterFLAP displayed reduced accuracy. WaterFLAP clearly placed the highest number of waters (98 versus 65 observed by crystallography), perhaps due to the large size of the BTK binding site.

Table 9. Number of placed water molecules by each program within the ‘area of interest’ of BTK (Figure 4b).

Compound	Crystal Structure	3D-RISM	SZMAP	WaterFLAP	WaterMap
6	7	9	10	11	7
7	8	10	8	11	8
8	6	10	10	11	6
9	6	10	9	12	6
10	7	9	8	11	8
11	9	8	10	10	8
12	7	9	9	11	9
13	9	8	14	11	9
14	6	7	9	10	6
Sum	65	80	87	98	67

Re-creation of the water network. For the water network analysis, the pairwise distances between the oxygen atoms of all eight key water molecules were calculated. The predicted water molecule distances were then compared to the crystallographic water network. Two diagonals were compared to capture the main distance relations between the water molecules (Figure 18a). Using only these two diagonals allowed the two water clusters (cluster 1: waters #0, #1, #2, and cluster 2: waters #4, #5, #6) to be compared, while at the same time the distances between the clusters and the single water molecules were captured, but the RMSD was not artificially lowered by including many large distances.

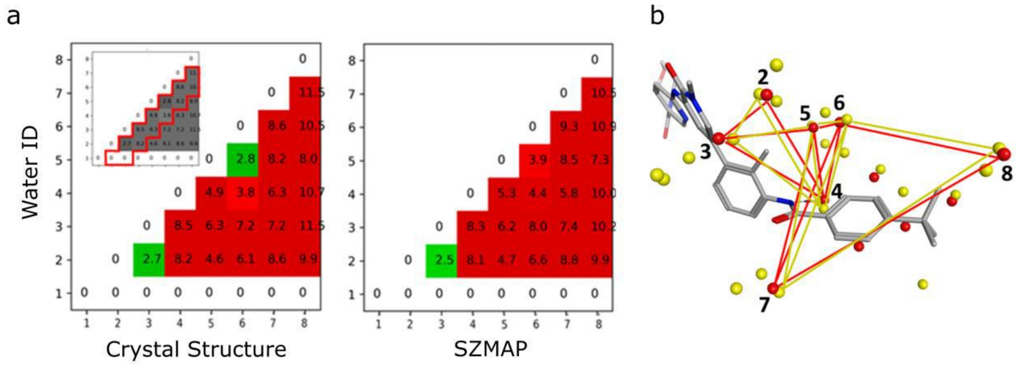


Figure 18. Re-creation of the water network within BTK complexes. (a) Pairwise distance map of crystallographically observed (left) and SZMAP placed (right) water molecules present in the BTK structure of **6**. Numbers in the matrix are the pairwise distances in Å between the oxygen atoms. Only the two highlighted diagonals were considered for comparison (inset in a). (b) Overlay of the crystallographically observed (red) and the SZMAP placed (yellow) water oxygens and network from this complex.

The crystallographically observed water clusters can be easily identified in the distance matrices (Figure 18a) because they are the only water molecules within hydrogen bond distances to each other. The distance matrices for the placed water molecules in the complex as well as holo version (Figures S12 and S13) were then compared to the crystallographically observed water distance matrix (Figure S11) by calculating the RMSDs (Table 10, Table S5).

WaterMap most accurately re-created the water network with an average RMSD of 0.48 Å (Table 10). Additionally, none of the placed water molecules had a greater distance than 2.5 Å from any crystallographically observed water molecule, i.e. all crystallographic waters were matched. WaterFLAP also re-created the water network well, with only one water molecule being missed. Due to the large number of placed water molecules by WaterFLAP and SZMAP (Table 9), the re-creation of the water network was less accurate. This ranking well reflected the correlation of placed and crystallographic waters (Table 9, Figure 16).

Table 10. RMSD of pair-wise distance matrices between crystallographic and computationally placed water molecules in the BTK protein-ligand structures. Numbers in parentheses (x/y) denote the number of placed water molecules x >2.5 Å away from X-ray water molecules y.

Compound	Crystal structure	3D-RISM		SZMAP		WaterFLAP		WaterMap	
6	0.0	0.45		0.70		3.09		0.31	
7	0.0	0.76		1.38	(1/6)	0.97		0.43	
8	0.0	0.69		1.31	(1/7)	0.62		0.18	
9	0.0	0.41		0.88		1.01		0.27	
10	0.0	0.56	(1/7)	0.80		0.94	(1/7)	0.50	
11	0.0	1.59	(1/8)	0.79		0.74		0.46	
12	0.0	0.98	(1/7)	0.93		0.62		1.09	
13	0.0	0.4		1.05		0.96		0.88	
14	0.0	0.73		0.84		0.4		0.2	
Averages		0.73	(3/65)	0.96	(2/65)	1.04	(1/65)	0.48	(0/65)

Energetic Contribution of Water Molecules and SAR Consistency. Four different aspects were analyzed concerning the prediction of water locations and energetics within BTK complexes: (1) Displacement of water #1; (2) Predicted energies for highly integrated waters #7 and #8; (3) Disruption of the cluster formed by waters #4, #5, #6; and (4) Average energy contribution for each water molecule. The displacement of water #1 was analyzed using the structure pairs **8/9** and **12/13** (Figure 19c and d). Structure pair **12/13** only differed by a methyl group attached to the pyridazinone ring – the methyl in **12** displaced water #1. In addition to the difference in the methyl group, the structure pair **8/9** also differed in its left-hand terminal group. In this region, the rings vary slightly, but since this portion points primarily into solvent, it was not expected to alter the predicted water molecule arrangement inside the 'area of interest'. For both structure pairs, the displacement of water #1 led to a decrease in affinity (from 0.0013 µM for **9** to 0.145 µM for **8** and from 0.0016 µM for **13** to 0.005 µM for **12**). Therefore, water #1 contributed favorably to the overall energy.

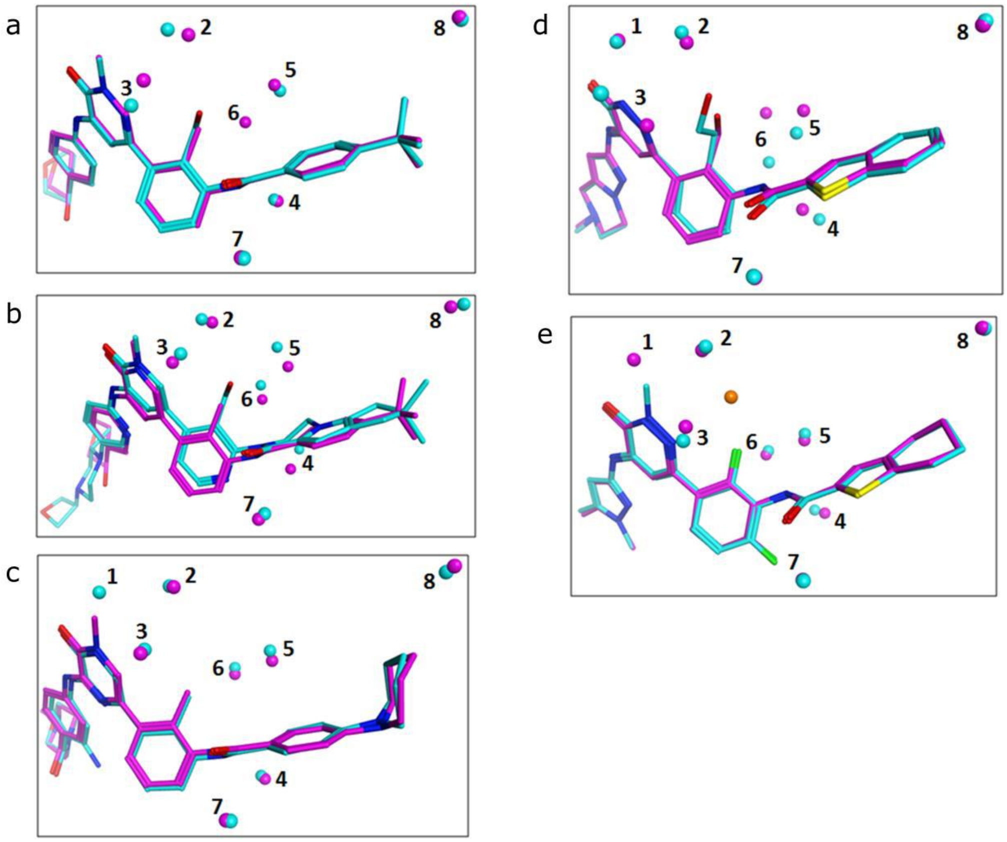


Figure 19. Selected structure pairs used for the analysis of the effects of ligand modifications on the water network. Ligand 2D diagrams are shown in Figure 2; their potencies are shown in Table 3. a) Structure pair **6/7** (magenta/cyan); b) **6**/**14** (magenta/cyan); c) **8/9** (magenta/cyan); d) **10/11** (cyan/magenta); e) **12/13** (cyan/magenta). Water oxygens are colored with respect to the ligand. In panel e), **12** contains an additional water molecule (orange). Synthesis of inhibitors can be found in Table S6.

Only SZMAP and WaterFLAP scored the predicted water #1 energetically favorably for both structure pairs (Table 11). According to 3D-RISM, water #1 in the **12/13** pair had nearly no energetic contribution, while in pair **8/9** it was predicted to be energetically unfavorable. WaterMap on the other hand rated water #1 in both cases as thermodynamically unstable, with an unfavorable positive energy score. WaterRank scores for the crystallographically observed water molecules were barely in the range of 'possible to replace'.

In addition to the favorable contribution of water #1, waters #7 and #8 were of interest due to their being buried in the binding pocket. Water #7 forms four interactions, one with the ligand, two with the protein and one with another water molecule. Water #8 displays three interactions with the protein. The high number of interactions between water #8 and protein coupled with the fact that this water could never be replaced by modifying the ligand (data not shown) indicated that waters #7 and #8 should receive favorable calculated energy scores. WaterFLAP most consistently scored waters #7 and #8 favorably in structure pair **12/13**, while SZMAP was most consistent for structure pair **8/9** (Table 11). 3D-RISM rated all waters in both structure pairs as favorably contributing, although the absolute energy values varied quite significantly (Table 11, 3D-RISM column). The predicted energies by SZMAP for water #8 in structure pair **12/13** were favorable for **12** and unfavorable for **13**. This might be due to the selection of the predicted waters, which is more difficult for SZMAP due to the higher number of predicted water molecules. WaterMap rated both waters in all four structures as highly unfavorable. In particular, water #7 received a very high energy score, which confirmed previous observations from BRD structures where a highly integrated water molecule also received a very unfavorable free energy score (□G > 5 kcal/mol). Overall, water #8 was in most cases scored more favorably than water #7. Finally, WaterRank scores for compounds **9**, **12**, and **13** were lower than expected for water molecules that displayed near ideal interaction geometries. Only **8** showed a high WaterRank score for water #8.

Table 11. Predicted free energies for water molecules of interest in the pairs of structures from Figure 19. WaterRank scores are for crystallographically observed water molecules, all other scores are for predicted waters.

Compound	Water ID	3D-RISM	SZMAP	WaterFLAP	WaterMap (ΔH)	WaterRank
6	2	-1.17	0.94	0.07	2.68 (-1.91)	3.8
6	3	-7.16	-7.05	-0.35	0.87 (-4.14)	2.2
6	4	4.99	-1.94	2.04	1.30 (-1.53)	0.0
6	5	-0.12	-4.55	-3.20	0.43 (-4.77)	3.9
6	6	5.03	-2.29	0.87	2.93 (-1.28)	2.3
7	3	-5.07	2.34	-1.05	2.30 (-2.53)	3.5
14	2	5.96	0.82	1.23	1.54 (-3.07)	1.0
14	3	-3.53	-6.60	0.76	2.33 (-2.48)	2.3
14	4	3.56	-1.76	-0.49	2.01 (0.28)	1.0
14	5	2.59	-7.15	-0.34	3.77 (-1.19)	2.2
14	6	0.96	-2.08	-0.14	4.77 (1.33)	2.3
10	4	5.52	-1.93	2.00	0.98 (-0.26)	0.0
10	5	4.61	-8.25	0.13	1.50 (-3.86)	2.3

10	6	0.14[a]	1.88	-0.67	3.12	(-0.19)	2.2
11	3	-5.45	-9.06	-0.68	2.60	(-1.79)	1.0
11	4	5.2	-4.20	-0.52	0.66	(-1.16)	0.0
11	5	6.21	-6.29	-1.78	2.69	(-2.63)	2.5
11	6	-1.51	1.89	2.68	4.14	(2.86)	2.3
9	1	1.54	-6.96	-3.40	6.99	(1.73)	2.3
9	7	-5.48	-3.19	3.13	8.52	(4.04)	2.2
9	8	-9.40	-7.88	-2.23	5.75	(0.65)	3.9
8	7	-3.13	-3.49	3.16	7.20	(2.77)	3.6
8	8	-11.19	-6.25	-1.77	5.64	(0.69)	5.6
13	1	0.25	-8.43	-2.66	2.99	(-2.00)	2.3
13	7	-4.12	-4.98	-1.25	7.02	(2.02)	2.2
13	8	-4.20	-5.92	-3.22	4.85	(-0.24)	3.9
12	7	-0.02	-6.63	-1.14	7.87	(2.93)	2.3
12	8	-8.90	0.75	-3.16	4.88	(-0.20)	3.9

a) The predicted water molecule was >2.5 Å away from the closest crystallographically observed one.

Three structure pairs – **6/7**, **6/14**, and **10/11** – were used to analyze the water cluster formed by waters #4, #5, #6 (Figure 19 a, b and e). Structure pair **6/7** differs by the presence of water #6 (Figure 19 a). However, the extension of this substituent from methyl to hydroxymethyl did not place the terminal hydroxy into the pocket where water is present. A closer examination of the temperature factor and electron density of water #6 in structure **7** suggests that its position may only be partially occupied. This led to the conclusion that the energetic contribution of water #6 should be neutral or unfavorable. Due to the greater hydrophilicity, this hydroxymethyl group when combined with an additional acceptor moiety due to the change from pyridine to pyridazine in the central linker ring stabilized the water network and resulted in an increased affinity of **7**. Water #3 is particularly stabilized due to a hydrogen bond interaction with the nitrogen of the pyridazine. Only WaterFLAP scored water #3 more favorably for **7** relative to **6** (Table 11). WaterScore also showed a difference in stabilization of water #3 (2.2 in **7** to 3.5 in **6**).

Despite the substantial structural differences between **6** and **14**, they align very well in the active site of BTK (Figure 19 b). Waters #4, #5, and #6 are shifted in the BTK/**14** complex due to cyclization relative to **6**. The change from the methyl to hydroxymethyl substitution helps stabilize water #2. Only SZMAP predicted all waters in **14** to have a negative energy score. WaterFLAP also predicted all three waters to be energetically favorable, however, the total energy contribution of the three was reduced relative to **6** with a significant loss in energy of water #5 (Table 11).

Structure pair **10/11** contains shifts of waters #3 and #6. As observed for the **6/7** pair, the shift of water #6 is not related to the extension of the hydroxymethyl substituent to a hydroxyethyl. Therefore, the energetic conclusions – water #6 being relatively neutral or slightly unfavorable – are supported. However, the substitution of the ligand disrupts the hydrogen bond of water #3 and the aromatic nitrogen of the pyridazine ring leads to a slight decrease in affinity (compare **11**, 0.0013µM to **10**, 0.0045µM). The programs all scored water #6 very differently, from favorable to unfavorable. SZMAP predicted scores for water #6 were the most consistent from **10** to **11** (Table 11).

The overall energy contribution of the different water molecules was then analyzed. Average scores for each water in the 9 BTK structures were calculated and compared among the different tools (Figure 20). Only very limited correlation for the predicted energies was observed for the different programs. The overall shapes of the curves (irrespective of the actual values) varied significantly. Interestingly, the average water scores predicted by WaterMap were all unfavorable, while the average SZMAP scores were all favorable.

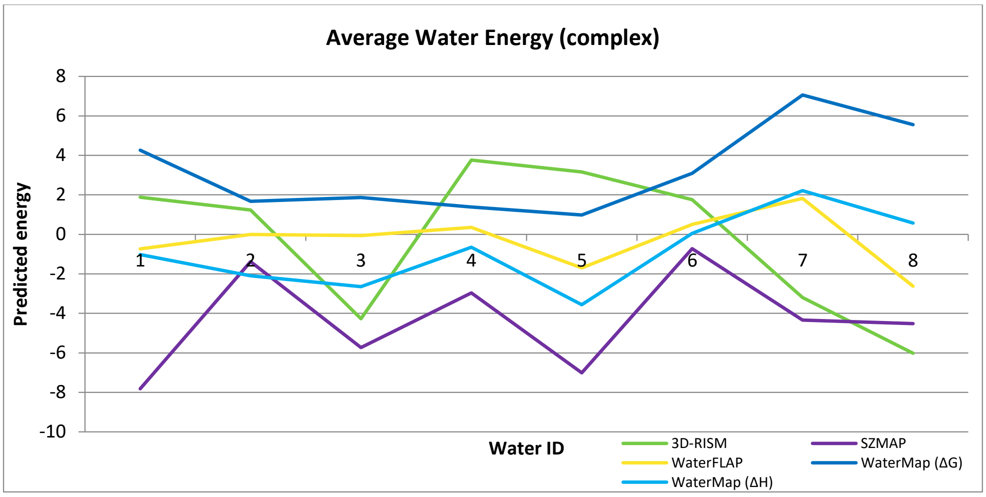


Figure 20. Average energies for predicted waters across all nine BTK complexes. Water ID is shown on the X-axis (middle of the plot). The total number of observed water molecules for each water position were as follows: #1, 4; #2, 9; #3, 8; #4, 9; #5, 9; #6, 8; #7, 9; and #8, 9. Connecting lines are for visualization purpose only.

Human Immunodeficiency Virus (HIV) Protease. As a last example, a tetrahedrally coordinated water (HOH-A-1037) from an HIV protease complex (1kzk[38]) was chosen for analysis. Due to the large number of hydrogen bonds this water forms to the protein and bound ligand, it was expected to be easy to predict as well as score.

Distance to Crystallographic *Water Molecules.* All tools only placed one water molecule in the HIV protease complex, which made the identification of the correct one very easy. Additionally, all placed water molecules were within 1 Å distance to the crystallographically observed one (Figure 21 and Table 12). WaterFLAP placed the water molecule most accurately (0.15 Å away from the crystallographically observed water). As observed previously, the accuracy of the water placement of 3D-RISM increased when the ligand was omitted.

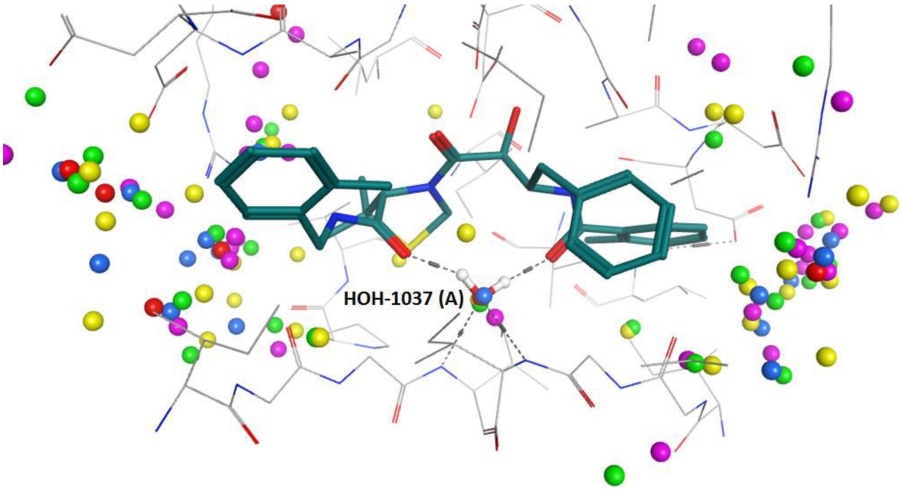


Figure 21. Placement of water oxygens in HIV protease (1kzk.pdb). The observed crystallographic water molecule of interest (HOH 1037 from chain A) is displayed in ball-and-stick (red oxygen with white hydrogens added by MOE). Other crystallographic water oxygens are in red. Water oxygens placed by WaterFLAP (yellow), SZMAP (purple), 3D-RISM (green), and WaterMap (blue) are superimposed.

Table 12. Distances (Å) between predicted and crystallographically observed water oxygens in the complex and holo form of HIV protease.

	3D-RISM	**SZMAP**	**WaterFLAP**	**WaterMap**
Complex	0.22	0.72	0.15	0.20
Holo	0.17	0.52	0.22	0.76

Re-creation of the water network. The observed crystallographic water molecule is interacting with four surrounding atoms, two on the ligand and two on the protein site. Interestingly, the hydrogen bond heavy atom distances to the protein are both longer than the ones to the ligand (Table 13). Analyzing the distances of the surrounding atoms to the predicted water molecules showed that both 3D-RISM and WaterFLAP shortened the distance to the nitrogen atom of Ile-B-50, while the other distances remain unchanged. SZMAP also shortened the distance to Ile-B-50, but at the same time the distances to the nitrogen of Ile-A-50 and O32 of the ligand JE2 increased to a non-optimal range. WaterMap kept the distances to the ligand atoms unchanged and shortened the distance to the nitrogen atom of Ile-A-50, but elongated the distance to the nitrogen atom of Ile-B-50.

Table 13. Distances between crystallographically observed (HOH-A-1037) and placed water oxygens to surrounding protein and ligand atoms in HIV protease (1kzk). Distances are measured between heavy atoms. Those in green indicate a more optimal hydrogen bond distance (2.6-2.9 Å), those in bold red a less optimal distance compared to the crystallographic ones.

Distance [Å]		Crystal structure	3D-RISM	SZMAP	WaterFLAP	WaterMap
Protein	Ile-A-50 N	3.00	3.02	2.51	3.09	2.89
	Ile-B-50 N	3.01	2.80	2.87	2.88	3.19
Ligand	JE2 O32	2.74	2.86	3.37	2.76	2.73
	JE2 O10	2.84	2.88	2.86	2.86	2.83

Energetic Contribution of Water Molecule and Consistency with SAR. This well-integrated water molecule was expected to contribute favorably to the binding affinity. Apart from being well integrated in the protein-ligand complex, this water has proven to be hard to displace and is present in the apo structure of HIV protease (Figure 22). Both SZMAP and WaterFLAP scored the crystallographically observed water molecule as thermodynamically stable, i.e. favorably contributing to ligand binding (Table 14). WaterRank on the other hand, which is a geometry-based score, rated the crystallographically observed water molecule to be between ‘easily’ and ‘possible’ to replace (Table 14, Figure 23). The relatively low score is likely related to the fact that only water-protein interactions are considered. Water-ligand interactions are not included in the geometry assessment made by WaterRank.

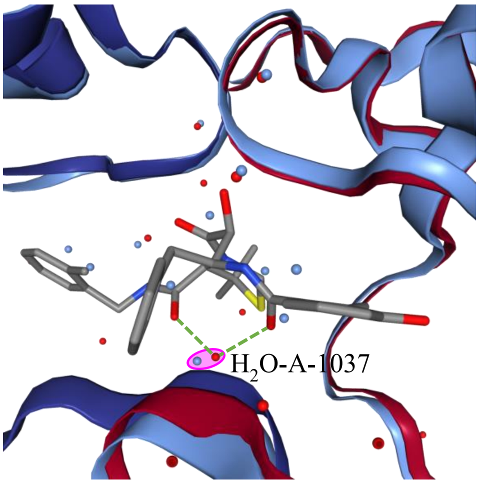


Figure 22. Superimposed HIV protease structures and water molecules. The apo structure of HIV protease (PDB accession code 1g6l) is shown in a blue ribbon with blue water oxygen atoms, superimposed onto the present ligand-bound complex (PDB accession code 1kzk) in a dark red ribbon with red water oxygens. Green dashed lines indicate hydrogen bonds between the ligand and water HOH-A-1037. The corresponding water molecule in the apo structure is enclosed by the pink circle.

Table 14. Predicted water scores for the crystallographically observed water molecule (HOH-A-1037) within the HIV protease complex (1kzk).

	SZMAP	WaterFLAP	WaterRank
Complex	-1.18	-4.69	2.2

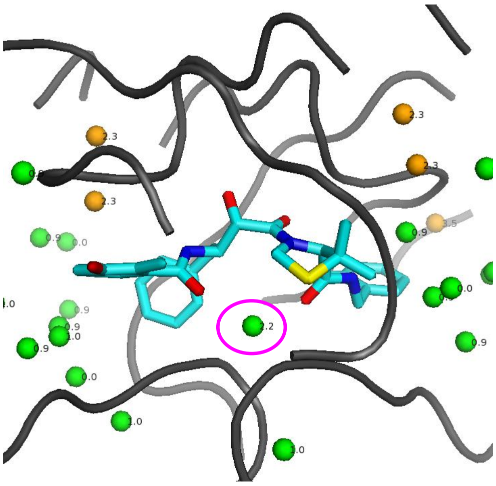


Figure 23. WaterRank scores for the HIV protease complex (1kzk) in the vicinity of the active site. The backbone ribbon of HIV protease is shown in grey, the ligand in light blue, and the WaterRank scores are color coded (‘easy to displace’ (scores of 0 – 2.3) in green, ‘possible to replace’ (scores of 2.3 – 4.0) in orange). The highly integrated water (HOH A 1037) is circled.

All water prediction tools except WaterMap rated the water molecule placed in the complex structure as thermodynamically stable (Table 15). 3D-RISM and SZMAP also scored the water molecule in the holo form as favorably contributing to ligand binding, while WaterFLAP predicted the water molecule to be slightly unfavorable. The WaterMap score increased for the placed water molecule in the holo structure, but it was still rated as thermodynamically unstable. As observed previously, the more restrained the water molecule is, the less favorable it was rated by WaterMap.

Table 15. Predicted water scores for placed water molecules in the HIV protease complex (1kzk).

	3D-RISM	SZMAP	WaterFLAP	WaterMap
Complex	-4.69	-5.68	-4.91	**5.01**
Holo	-5.68	-2.99	**0.16**	**1.27**

Conclusions.

Four different water prediction tools – 3D-RISM, SZMAP, WaterFLAP, and WaterMap – were analyzed for their abilities to accurately predict water molecule locations and energetic contributions. Four different aspects relevant for a qualitative assessment of the tools were analyzed – accuracy in water placement, number of predicted water molecules, re-creation of the hydrogen bond network, and correlation of water energies with observed SAR. WaterRank scores were also calculated as a null hypothesis test, since these scores are based solely on an analysis of the geometric coordination of each crystallographic water molecule. To support these analyses, multiple new protein/ligand crystallographic complexes (2 BRD and 7 BTK) were determined at high resolution. These data, combined with previously published data from these proteins as well as an HIV protease formed the test set for assessing water energy prediction software.

Overall, the placement of water molecules was fairly accurate, with all programs predicting 60-90% of water oxygens within 1 Å of their observed locations. WaterMap in particular achieved very high accuracies for all analyzed systems – BRD, BTK, and HIV, with >=70% of water oxygens located within 0.5 Å of observed. The other tools achieved somewhat reduced placement accuracies, but at the same time their results were less easy to interpret due to the higher number of predicted water oxygens resulting from the calculations. This is to some extent due to the original purpose of the tools. For example, SZMAP was not developed to predict specific water molecule locations, but rather to generate ideas for potential alterations of the ligand. Therefore, more waters are placed than would actually fit into the binding pocket.

The results for the re-creation of the water network were variable for the protein families analyzed. For BRD compounds, WaterFLAP and WaterMap achieved almost the same accuracies. However, WaterMap was not able to accurately place 4 out of 41 crystallographically observed waters while WaterFLAP found all except one. In contrast, the re-creation of the water network in BTK structures was most accurate for WaterMap, which did not miss any crystallographic waters. By contrast, WaterFLAP was significantly less accurate for BTK, missing only one out of 66 waters.

A major problem for all tools was a consistent prediction of energetic contributions of water molecules. The tools seldom agreed with each other and also the consistency within each tool was very low. WaterMap showed specific drawbacks when highly integrated and constrained water molecules were evaluated. Without exception, WaterMap scored them unfavorably, even when the experimental SAR suggested otherwise. WaterFLAP scores on the other hand seemed to be strongly dependent on the number of hydrogen bonds potentially formed by the water molecules. The scoring of crystallographically observed water molecules led to more stable results, however, the different methods did not agree with each other. For BRD structures, WaterFLAP and WaterRank achieved similar rankings of the waters of interest. Even for the rather easy example of water molecule A-1037 within HIV protease, the predicted energies were not consistent. 3D-RISM, SZMAP, and WaterFLAP all scored the water molecule highly favorable, while WaterMap predicted it to be unfavorable. Interestingly, according to WaterRank, the water molecule is between ‘easy and ‘possible’ to replace. This observation is certainly contrary to the visual observation of the tight 4-coordinate coordination of this water.

Overall, diverse methods from the straightforward WaterRank, to the simulation heavy WaterMap, were analyzed throughout this study using different protein systems – Bromodomains, Bruton's Tyrosine Kinases, and an HIV protease – with no clear advantage of one tool over any other emerging. Based on these results, we recommend that future development of water prediction tools should focus specifically on a more consistent score prediction. Small variations in the water position should not have a dramatic effect on the predicted energy. This would lead to a more accurate and more reliable prediction, which is important if these tools are to be used for prospective predictions of potencies when deciding on which water to target for replacement.

One can argue that the shifts in ligands from structure to structure can affect these calculations, particularly when one is attempting to equate calculated energetic differences of displaced waters to observed differences in potencies. Changes in ligand structure also change the nature of direct protein-to-ligand contacts, which can also influence binding affinity and cloud these conclusions. We have attempted to minimize these effects by carefully choosing pairs of ligands with very similar binding modes and virtually superimposable protein active sites. We hoped that we would be able to discern relative differences within chemical series, perhaps rank ordering ligands from more to less potent based on water energetics, as opposed to making absolute predictions of potency differences. This proved not to be the case.

As previously described, all input protein/ligand/water complexes were initially processed using the recommended protein preparation procedure within each program. Thus, each complex was set up (histidine tautomers, Asn and Gln flips, protonation states, etc.) with crystallographically observed waters present. This provided a 'best case' scenario when waters and ligand were removed for prediction by the programs. One could argue that, for a true test of a program's ability to place waters accurately, each complex should be re-prepared without waters and ligand present prior to the water prediction step. This was not done in the present work.

So what can medicinal chemists take away from this analysis? The good news is that most crystallographic waters can be predicted by these programs with a few notable exceptions. The bad news is that energetics calculated by water prediction software should not be used for compound design. One will simply have to target key predicted waters with ligand modifications to assess impact on potency. Fortunately, given the reasonable accuracy in predictions of crystallographically observed water locations, one can use these calculations as a guide for where to substitute ligands to affect crystallographic water displacement. These results will be furnished to the vendors, with the hope that future refinements in the energy calculations will yield more robust correlations with observed changes in potency.

ASSOCIATED CONTENT

Supporting Information.

Additional tables and figures (PDF).

Table with calculated energies for the crystallographic observed and placed water molecules for BRD and BTK compounds (XLS).

Molecular formula strings (CSV).

AUTHOR INFORMATION

Corresponding Authors

* Paul Gibbons: gibbons.paul@gene.com

* Eva Nittinger: nittinger@zbh.uni-hamburg.de

* Charles Eigenbrot: charleseigenbrot@comcast.net

Present Addresses

† Current Address: Gilead Sciences Inc., 333 Lakeside Drive, Foster City, CA 94404 USA

†† Current Address: Relay Therapeutics, 215 First Street, 3rd Floor, Cambridge, MA 02142

Author Contributions

EN hast written the manuscript, developed and conducted the evaluation strategy. DD, CE, JK, JM, and YT determined the BRD and BTK crystal structures used in the analysis. VT has supervised the project. DFO and PG have contributed to the manuscript and have supervised the project.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Matthias Rarey for his support during the project. We thank Terry Crawford, Shumei Wang, Lina Chan, Alex Cote, Chris Nasveschuk, and Matthew Berlin for their syntheses of BRD and TAF small molecule inhibitors. We also acknowledge Wendy Young, Gina Wang, Kevin Currie, and the other chemists at CGI Pharmaceuticals (now Gilead) for their support of the BTK program and syntheses of BTK small molecule inhibitors. Additionally, we thank Laura E. Zawadzke and Eneida Pardo of Constellation Pharmaceuticals for their help assaying the BRD and TAF inhibitors; and Julie DiPaolo for conducting the BTK Lanthascreen assays.

Results shown in this report are derived from work performed at Argonne National Laboratory, Structural Biology Center (SBC) at the Advanced Photon Source. SBC-CAT is operated by UChicago Argonne, LLC, for the U.S. Department of Energy, Office of Biological and Environmental Research under contract DE-AC02-06CH11357. Use of the

Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research, and by the National Institutes of Health, National Institute of General Medical Sciences (including P41GM103393) . The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS or NIH. The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health, National Institute of General Medical Sciences, and the Howard Hughes Medical Institute. The Advanced Light Source is a Department of Energy Office of Science User Facility under Contract No. DE-AC02-05CH11231.

ABBREVIATIONS

BRD, Bromodomain; BTK, Bruton's Tyrosine Kinase; HIV, Human Immunodeficiency Virus.

REFERENCES

1. Abel, R., Young, T., Farid, R., Berne, B. J., & Friesner, R. A. (2008). Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. Journal of the American Chemical Society, 130(9), 2817–2831. doi:10.1021/ja0771033
2. Lazaridis, T. (1998). Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. The Journal of Physical Chemistry B, 102(18), 3531–3541. doi:10.1021/jp9723574
3. Lazaridis, T. (1998). Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids. The Journal of Physical Chemistry B, 102(18), 3542–3550. doi:10.1021/jp972358w
4. Graham, S. E., Smith, R. D., & Carlson, H. A. (2018). Predicting Displaceable Water Sites Using Mixed-Solvent Molecular Dynamics. Journal of Chemical Information and Modeling, acs.jcim.7b00268. doi:10.1021/acs.jcim.7b00268
5. Adams, D. J. (1975). Grand canonical ensemble Monte Carlo for a Lennard-Jones fluid. Molecular Physics, 29(1), 307–311. doi:10.1080/00268977500100221
6. Michel, J., Tirado-Rives, J., & Jorgensen, W. L. (2009). Prediction of the water content in protein binding sites. Journal of Physical Chemistry B, 113(40), 13337–13346. doi:10.1021/jp9047456
7. Barillari, C., Taylor, J., Viner, R., & Essex, J. W. (2007). Classification of water molecules in protein binding sites. Journal of the American Chemical Society, 129(9), 2577–2587. doi:10.1021/ja066980q
8. Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F., & Mason, J. S. (2007). A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): Theory and application. Journal of Chemical Information and Modeling, 47(2), 279–294. doi:10.1021/ci600253e
9. Bayden, A. S., Moustakas, D. T., Joseph-McCarthy, D., & Lamb, M. L. (2015). Evaluating Free Energies of Binding and Conservation of Crystallographic Waters Using SZMAP. Journal of Chemical Information and Modeling, 55(8), 1552–1565. doi:10.1021/ci500746d
10. Kovalenko, A., & Hirata, F. (1999). Self-consistent description of a metal--water interface by the Kohn--Sham density functional theory and the three-dimensional reference interaction site model. The Journal of Chemical Physics, 110(20), 10095–10112. doi:10.1063/1.478883
11. Kovalenko, A., & Hirata, F. (1998). Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach. Chemical Physics Letters, 290(1–3), 237–244. doi:10.1016/S0009-2614(98)00471-0
12. Nguyen, C. N., Young, T. K., & Gilson, M. K. (2012). Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. Journal of Chemical Physics, 137(14), 044101. doi:10.1063/1.4751113
13. Kellogg, G. E., & Chen, D. L. (2004). The importance of being exhaustive. Optimization of bridging structural water molecules and water networks in models of biological systems. Chemistry and Biodiversity, 1(1), 98–105. doi:10.1002/cbdv.200490016
14. Amadasi, A., Surface, J. A., Spyrakis, F., Cozzini, P., Mozzarelli, A., & Kellogg, G. E. (2008). Robust

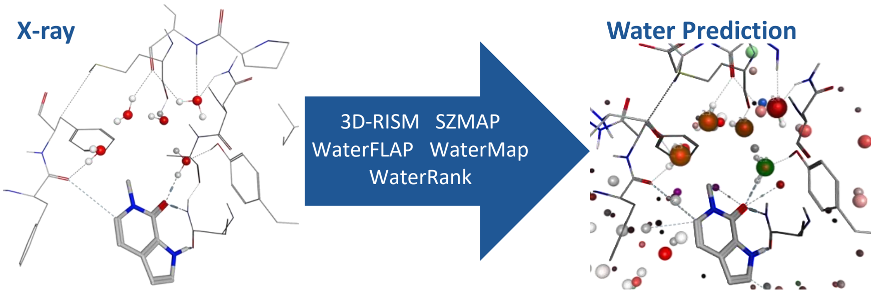
classification of “relevant” water molecules in putative protein binding sites. Journal of Medicinal Chemistry, 51(4), 1063–1067. doi:10.1021/jm701023h

15. Pitt, W. R., & Goodfellow, J. M. (1991). Modelling of solvent positions around polar groups in proteins. Protein Engineering, Design and Selection, 4(5), 531–537. doi:10.1093/protein/4.5.531
16. Pitt, W. R., Murray-Rust, J., & Goodfellow, J. M. (1993). AQUARIUS2: Knowledge-based modeling of solvent sites around proteins. Journal of Computational Chemistry, 14(9), 1007–1018. doi:10.1002/jcc.540140902
17. Verdonk, M. L., Cole, J. C., & Taylor, R. (1999). SuperStar: A Knowledge-based Approach for Identifying Interaction Sites in Proteins. Journal of Molecular Biology, 289(4), 1093–1108. doi:10.1006/jmbi.1999.2809
18. Verdonk, M. L., Cole, J. C., Watson, P., Gillet, V., & Willett, P. (2001). Superstar: improved knowledge-based interaction fields for protein binding sites11Edited by R. Huber. Journal of Molecular Biology, 307(3), 841–859. doi:10.1006/jmbi.2001.4452
19. Raymer, M. L., Sanschagrin, P. C., Punch, W. F., Venkataraman, S., Goodman, E. D., & Kuhn, L. a. (1997). Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. Journal of molecular biology, 265(4), 445–64. doi:10.1006/jmbi.1996.0746
20. Beuming, T., Farid, R., & Sherman, W. (2009). High-energy water sites determine peptide binding affinity and specificity of PDZ domains. Protein Science, 18(8), 1609–1619. doi:10.1002/pro.177
21. Chrencik, J. E., Patny, A., Leung, I. K., Korniski, B., Emmons, T. L., Hall, T., … Benson, T. E. (2010). Structural and Thermodynamic Characterization of the TYK2 and JAK3 Kinase Domains in Complex with CP-690550 and CMP-6. Journal of Molecular Biology, 400(3), 413–433. doi:10.1016/j.jmb.2010.05.020
22. Laha, J. K., Zhang, X., Qiao, L., Liu, M., Chatterjee, S., Robinson, S., … Cuny, G. D. (2011). Structure-activity relationship study of 2,4-diaminothiazoles as Cdk5/p25 kinase inhibitors. Bioorganic and Medicinal Chemistry Letters, 21(7), 2098–2101. doi:10.1016/j.bmcl.2011.01.140
23. Repasky, M. P., Murphy, R. B., Banks, J. L., Greenwood, J. R., Tubert-Brohman, I., Bhat, S., & Friesner, R. A. (2012). Docking performance of the glide program as evaluated on the Astex and DUD datasets: A complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. Journal of Computer-Aided Molecular Design, 26(6), 787–799. doi:10.1007/s10822-012-9575-9
24. Knegtel, R. M. A., & Robinson, D. D. (2011). A role for hydration in interleukin-2 inducible T cell kinase (Itk) selectivity. Molecular Informatics, 30(11–12), 950–959. doi:10.1002/minf.201100086
25. Nguyen, C. N., Cruz, A., Gilson, M. K., & Kurtzman, T. (2014). Thermodynamics of Water in an Enzyme Active Site : Grid-Based Hydration Analysis of Coagulation Factor Xa. Journal of Chemical Theory and Computation, 10(7), 2769–2780.
26. Bodnarchuk, M. S., Viner, R., Michel, J., & Essex, J. W. (2014). Strategies to Calculate Water Binding Free Energies in Protein − Ligand Complexes. Journal of chemical information and modeling, (54), 1623–1633.
27. Mason, J. S., Bortolato, A., Congreve, M., & Marshall, F. H. (2012). New insights from structural biology into the druggability of G protein-coupled receptors. Trends in Pharmacological Sciences, 33(5), 249–260. doi:10.1016/j.tips.2012.02.005
28. Bortolato, A., Tehan, B. G., Bodnarchuk, M. S., Essex, J. W., & Mason, J. S. (2013). Water network perturbation in ligand binding: Adenosine A2A antagonists as a case study. Journal of Chemical Information and Modeling, 53(7), 1700–1713. doi:10.1021/ci4001458
29. Spyrakis, F., Ahmed, M. H., Bayden, A. S., Cozzini, P., Mozzarelli, A., & Kellogg, G. E. (2017). The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. Journal of Medicinal Chemistry, 60(16), 6781–6827. doi:10.1021/acs.jmedchem.7b00057
30. Crawford, T. D., Tsui, V., Flynn, E. M., Wang, S., Taylor, A. M., Côté, A., … Cochran, A. G. (2016). Diving into the Water: Inducible Binding Conformations for BRD4, TAF1(2), BRD9, and CECR2 Bromodomains. Journal of Medicinal Chemistry, 59(11), 5391–5402. doi:10.1021/acs.jmedchem.6b00264
31. Albrecht, B. K., Gehling, V. S., Hewitt, M. C., Vaswani, R. G., Côté, A., Leblanc, Y., … Audia, J. E. (2016). Identification of a Benzoisoxazoloazepine Inhibitor (CPI-0610) of the Bromodomain and Extra-Terminal (BET) Family as a Candidate for Human Clinical Trials. Journal of Medicinal Chemistry, 59(4), 1330–1339. doi:10.1021/acs.jmedchem.5b01882
32. Johnson, A. R., Kohli, P. B., Katewa, A., Gogol, E., Belmont, L. D., Choy, R., … Young, W. B. (2016). Battling Btk Mutants with Noncovalent Inhibitors That Overcome Cys481 and Thr474 Mutations. ACS Chemical Biology, 11(10), 2897–2907. doi:10.1021/acschembio.6b00480
33. Di Paolo, J. A., Huang, T., Balazs, M., Barbosa, J., Barck, K. H., Bravo, B. J., … Currie, K. S. (2011). Specific Btk inhibition suppresses B cell– and myeloid cell–mediated arthritis. Nature Chemical Biology, 7(1), 41–50. doi:10.1038/nchembio.481
34. Young, W. B., Barbosa, J., Blomgren, P., Bremer, M. C., Crawford, J. J., Dambach, D., … Currie, K. S. (2016). Discovery of highly potent and selective Bruton’s tyrosine kinase inhibitors: Pyridazinone analogs

with improved metabolic stability. Bioorganic and Medicinal Chemistry Letters, 26(2), 575–579. doi:10.1016/j.bmcl.2015.11.076

35. Young, W. B., Barbosa, J., Blomgren, P., Bremer, M. C., Crawford, J. J., Dambach, D., … Currie, K. S. (2015). Potent and selective Bruton's tyrosine kinase inhibitors: Discovery of GDC-0834. Bioorganic & Medicinal Chemistry Letters, 25(6), 1333–1337. doi:10.1016/j.bmcl.2015.01.032
36. Wang, X., Barbosa, J., Blomgren, P., Bremer, M. C., Chen, J., Crawford, J. J., … Young, W. B. (2017). Discovery of Potent and Selective Tricyclic Inhibitors of Bruton's Tyrosine Kinase with Improved Druglike Properties. ACS Medicinal Chemistry Letters, 8(6), 608–613. doi:10.1021/acsmedchemlett.7b00103
37. Crawford, J. J., Johnson, A. R., Misner, D. L., Belmont, L. D., Castanedo, G., Choy, R., … Young, W. B. (2018). Discovery of GDC-0853: A Potent, Selective, and Noncovalent Bruton's Tyrosine Kinase Inhibitor in Early Clinical Development. Journal of Medicinal Chemistry, 61(6), 2227–2245. doi:10.1021/acs.jmedchem.7b01712
38. Reiling, K. K., Endres, N. F., Dauber, D. S., Craik, C. S., & Stroud, R. M. (2002). Anisotropic dynamics of the JE-2147-HIV protease complex: Drug resistance and thermodynamic binding mode examined in a 1.09 ?? structure. Biochemistry, 41(14), 4582–4594. doi:10.1021/bi011781z
39. Huang, P. P., Randolph, J. T., Klein, L. L., Vasavanonda, S., Dekhtyar, T., Stoll, V. S., & Kempf, D. J. (2004). Synthesis and antiviral activity of P1′ arylsulfonamide azacyclic urea HIV protease inhibitors. Bioorganic and Medicinal Chemistry Letters, 14(15), 4075–4078. doi:10.1016/j.bmcl.2004.05.036
40. Nittinger, E., Schneider, N., Lange, G., & Rarey, M. (2015). Evidence of water molecules--a statistical evaluation of water molecules based on electron density. Journal of chemical information and modeling, 55(4), 771–83. doi:10.1021/ci500662d
41. Meyder, A., Nittinger, E., Lange, G., Klein, R., & Rarey, M. (2017). Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. Journal of Chemical Information and Modeling, 57(10), 2437–2447. doi:10.1021/acs.jcim.7b00391
42. Bietz, S., Inhester, T., Lauck, F., Sommer, K., von Behren, M. M., Fährrolfes, R., … Rarey, M. (2017). From cheminformatics to structure-based design: Web services and desktop applications based on the NAOMI library. Journal of Biotechnology, 261, 207–214. doi:10.1016/j.jbiotec.2017.06.004
43. WaterRank – Desert Scientific Software. (n.d.).
44. Bietz, S., & Rarey, M. (2015). ASCONA: Rapid Detection and Alignment of Protein Binding Site Conformations. Journal of Chemical Information and Modeling, 55(8), 1747–1756. doi:10.1021/acs.jcim.5b00210
45. Bietz, S., & Rarey, M. (2016). SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles. Journal of Chemical Information and Modeling, 56(1), 248–259. doi:10.1021/acs.jcim.5b00588

TOC.



Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, den 21. März 2018

Eva Nittinger