Improving Rapid Affinity Calculations for Drug-Protein Interactions

by

Gregory Antonio Ross

St. Anne's College
and
Department of Biochemistry

Trinity 2013

A thesis submitted in partial fulfillment for the degree of
Doctor of Philosophy at the University of Oxford

Abstract

Improving Rapid Affinity Calculations for Drug-Protein Interactions

by Gregory Antonio Ross,
St. Anne's College, Oxford.

Submitted for the degree of Doctor of Philosophy, Trinity 2013

The rationalisation of drug potency using three-dimensional structures of protein-ligand complexes is a central paradigm in medicinal research. For over two decades, a major goal has been to find the rules that accurately relate the structure of any protein-ligand complex to its affinity. Addressing this problem is of great concern to the pharmaceutical industry, which uses virtual screens to computationally assay up to many millions of compounds against a protein target. A fast and trustworthy affinity estimator could potentially streamline the drug discovery process, reducing reliance on expensive wet lab experiments, speeding up the discovery of new hits and aiding lead optimization.

Water plays a critical role in drug-protein interactions. To address the often ambiguous nature of water in binding sites, a water placement method was developed and found to be in good agreement with X-ray crystallography, neutron diffraction data and molecular dynamics simulations. The method is fast and has facilitated a large scale study of the statistics of water in ligand binding sites, as well as the creation of models pertaining to water binding free energies and displacement propensities, which are of particular interest to medicinal chemistry.

Structure-based scoring functions employing the explicit water models were developed. Surprisingly, these attempts were no more accurate than the current state of the art, and the models suffered from the same inadequacies which have plagued all previous scoring functions. This suggests a unifying cause behind scoring function inaccuracy. Accordingly, mathematical analyses on the fundamental uncertainties in structure-based modelling were conducted. Using statistical learning theory and information theory, the existence of inherent errors in empirical scoring functions was proven. Among other results, it was found that even the very best generalised structure-based model is significantly limited in its accuracy, and protein-specific models are always likely to be better. The theoretical framework developed herein hints at modelling strategies that operate at the leading edge of achievable accuracy.

Acknowledgements

I am greatly indebted to my supervisors Philip Biggin and Garrett Morris for all their advice, suggestions, and trust, which guided my meandering path through computational chemistry and biochemistry. This project springs from their thoughts on developing "wet pharmacophore" models. I am grateful to Phil for always having an open door and ear, even for my less than bright ideas.

I have been very lucky to have been a part of the Structural Bioinformatics and Computational Biochemistry Unit. For the good times and work times, warm thanks to past and present members, particularly Amanda Buyan, Sonya Hanson, Joseph Goose, Heidi Koldsø, Joanna Lee, Craig Lumb, Jerome Ma, Maria Musgaard, Benjamin Morris, Michelle Sahai, David Shorthouse, Phillip Stansfeld, Matthias Schmidt, Lukas Stelzl, Jemma Trick, Ranjit Vijayan and Shabana Vohra. A special mention goes to Lukas and Sonya for our thermodynamic forays (LLAS), and Dave and Jemma for lunchtime mirth.

I am proud to have been accepted into the Doctoral Training Centre, and grateful to have emerged from it an enlightened and confused interdisciplinarian. I was in fact accepted onto a brand new course I had previously never heard of: "Systems Approaches to Biomedical Science Industrial Doctorate Centre", and I am glad that I was. I must acknowledge my fellow IDCers, especially for our evening scientific meet-ups throughout my DPhil. Thanks and sorry to Timothy Rooney and Konrad Krawczyk for their interesting discussions, and for the collaborations that never happened.

I am especially grateful to have met Jessica McGillen early in my DPhil. Thank you Jess, for your patience and insight when discussing my science-related ramblings, and for all time you have spent checking over my often poorly written write-ups. Above all else, you have turned my already happy time in Oxford into something truly special. My mum, dad and brother have been incredibly supportive of my prolonged studies, and I have always counted on their help and endless encouragement. Mum and Dad, this thesis is for you.

I am thankful to the Doctoral Training Centre and InhibOx for their funding, without which this research would not exist. I can only hope that their investment in me has been, and will be, worthwhile to people other than myself.

List of Publications Originating from this Thesis

Ross, G. A., Morris, G. M., Biggin, P. C., "Rapid and accurate prediction and scoring of water molecules in protein binding sites." *PLoS ONE* 7 (3), e32036+, Mar. 2012.

Ross, G. A., Morris, G. M., Biggin, P. C., "One size does not fit all: The limits of structure-based models in drug discovery.," *J. Chem. Theory Comput.* 9 (9), 4266-4274, Aug. 2013.

Ross, G. A., Morris, G. M., Biggin, P. C., "Explicit water in scoring functions: a bias-variance tradeoff.," *In preparation.*

Contents

List of Figures

List of Tables

Frequently used Abbreviations

AIC	**A**kaike's **I**nformation **C**riterion
AUC	**A**rea **U**nder **C**urve
CSAR	**C**ommunity **S**tructure **A**ctivity **R**esource
GBSA	**G**eneralised **B**orn **S**urface **A**rea
GBT	**G**radient **B**oosted **T**ree
HIV	**H**uman **I**mmunodeficiency **V**irus
IFST	**I**nhomogeneous **F**luid **S**olvation **T**heory
ITC	**I**sothermal **T**itration **C**alorimetry
JAWS	**J**ust **A**dd **W**ater**S**
MAE	**M**ean **A**bsolute **E**rror
MD	**M**olecular **D**ynamics
MMSE	**M**inimum **M**ean **S**quared **E**rror
MSE	**M**ean **S**quared **E**rror
NMR	**N**uclear **M**agnetic **R**esonance
PBSA	**P**oisson-**B**oltzmann **S**urface **A**rea
PDB	**P**rotein **D**ata **B**ank
PDF	**P**robability **D**istribution **F**unction
PMF	**P**otential of **M**ean **F**orce
QSAR	**Q**uantitative **S**tructure **A**ctivity Relationship
RISM	**R**eference **I**nteraction **S**ite **M**odel
RMSD	**R**oot **M**ean **S**quared **D**eviation
ROC	**R**eciever **O**perating **C**haracteristic

For my parents

Note that the problem is not to find the ρ_0 which correctly describes the "true physical situation". That is unknown, and always remains so, because of incomplete information. In order to have a usable theory we ask the much more modest question: "What ρ_0 best describes our state of knowledge about the physical situation".

Edwin Thompson Jaynes

Chapter 1

Introduction

1.1 Where are we now?

Drug discovery is difficult. The entire process, from selecting the target to clinical trials, can take over a decade and cost over a billion pounds[1]. Modern medical successes demonstrate how much has been achieved while simultaneously highlighting the many complex diseases that remain untreated. When pharmaceutical companies seek to improve on drugs already on the market, new compounds not only have to combat the disease they were designed to treat, but also must compete with existing drugs. These factors have contributed to difficulties faced by the pharmaceutical industry in recent years: research development costs are rising but productivity is falling[1–3]. This situation demands that our scientific understanding of drugs and proteins evolves in step with this ever-ratcheting challenge.

Computational methods form an important component of drug discovery[4]. Yet despite being over thirty decades old, the field of computer-aided drug design is still in its infancy, and there remain many hurdles to overcome before the dream of cheaper and more efficient drug development can be fulfilled[5–7].

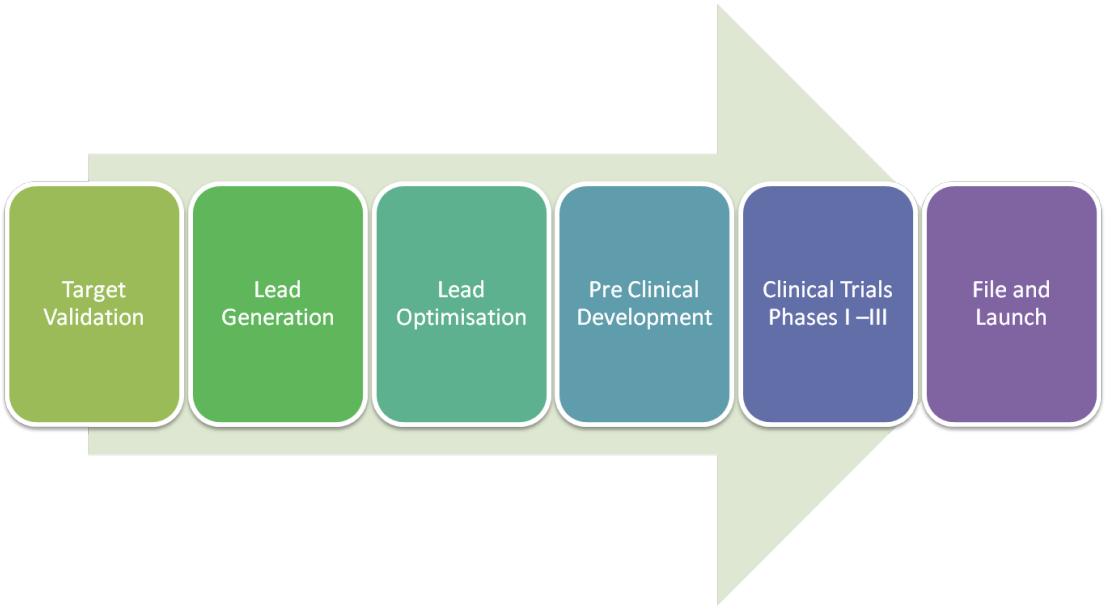


FIGURE 1.1: A representation of the main phases in drug discovery projects[1,8]. During the lead generation phase, many millions of compounds are assayed in high throughput screens. Later, in lead optimisation, chemical modifications are made to compounds in order to improve the interaction with the protein. Computational methods are playing increasingly important roles in these phases in predicting the potency of compounds against the protein target[4].

1.2 Strategies in Drug Discovery

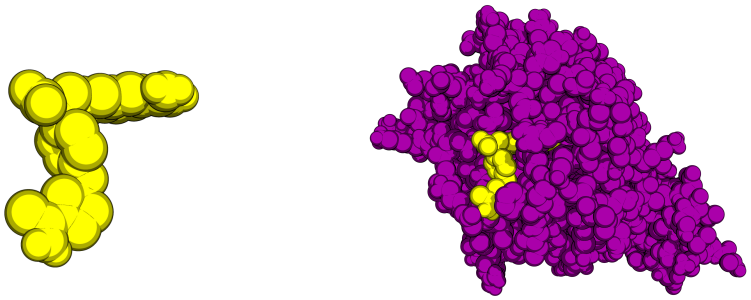
Whether a drug is designed to activate, or inactivate a protein, or inhibit the binding of endogenous ligands, it is fundamental that the drug and protein associate with sufficient strength. The degree to which two molecules bind is known as the affinity.

Early in a drug discovery project, high throughput chemical screens are used to determine compounds, known as "hits", which bind to the target protein with enough affinity or activity to merit further development[9,10]. The emergence of fully automated procedures in the 1980s was a major leap forward for the pharmaceutical industry, allowing vast libraries of compounds to be assayed without prior knowledge of the required ligand chemistry[9]. While modern techniques have the remarkable capacity to screen over 100,000 molecules a day[11], high throughput screens are expensive and require highly specialised laboratories. As a result, the increasing availability of cheap and powerful

computers has enabled so called "virtual" screens to become well established in the pharmaceutical and biotechnology industries, as well as in academic drug discovery[12]. Often, they are used to enrich libraries prior to high throughput screens with compounds that are more likely to be hits. In less costly projects, the very top hits from a virtual screen are taken forward for manual experimentation. The testing of up to many millions of compounds requires that only simple models be used to estimate affinity, so that the calculations can be carried out as fast as possible.

Once promising leads for drug candidates (unsurprisingly called "leads") have been identified, efforts focus on improving the affinity, activity and selectivity of the ligands. Early considerations can also be made with respect to the toxicity of the compounds, and whether they will be well-tolerated by the body[13]. Theoretical models play a crucial role in determining which chemical modifications will be pursued to enhance the affinity or activity of a compound. Most commonly, these models are qualitative structure-activity relationship (SAR) models, which are formal hypotheses regarding the relationship between ligand chemical structure and activity or affinity. Quantitative features of the leads, such as lipophilicity and charge, can also form the basis of mathematical models called quantitative structure-activity relationship (QSAR) models. As only a handful of compounds may reach this stage, it is also feasible to apply computational techniques that are more intensive than those used in virtual screening.

Determination of the three-dimensional structure of the target protein opens the possibility to powerful rationalisations of drug affinity that are not possible with only ligand chemical data[14]. With structural protein data, medicinal chemists try to optimise compounds to have favourable interactions with the protein[15]. One of the most effective strategies involves designing ligands to fit the three-dimensional shape of the binding site and to complement the hydrophobic, electrostatic, and hydrogen bonding surface[16,17].



(A) (B)

FIGURE 1.2: Drugs are designed to complement the shape of the target protein's binding site. (a) An example of a drug for the blood coagulation protein, factor Xa, demonstrating the characteristic hinge, like an upside-down "L", which is typical to many factor Xa inhibitors. (b) The complex of the protein (in purple) and ligand (in yellow), showing the drug is well buried in the binding site. Subfigures (a) and (b) are in different scales.

Implicit in structure-based lead optimisation strategies is the hypothesis that the three-dimensional structure of a protein-ligand complex – and, in particular, the pose of the ligand within the binding site – encodes the ligand's potency. Critically, the development of an accurate and generally applicable structure-based affinity prediction technique remains one of the most significant unsolved problems in drug discovery[18], and is subject to intense and continuous research. With such a technique, medicinal chemists would be well informed as to which chemical modifications to pursue. A *fast* and accurate affinity model would be of even greater benefit, not only for convenience in lead optimisation, but to greatly improve the productivity of virtual screens.

1.3 Computational approaches to binding affinity prediction

In the absence of structural information for a protein or protein-ligand complex, the chemical make-up of a known ligand provides a sufficient starting point for a virtual screen. Chemically similar ligands often have similar affinities and activities[19], so given

a ligand with a desirable potency against the target, a virtual library of compounds can be trawled to find compounds that match a particular description of that ligand. Descriptions can include chemical composition[20] or estimated three-dimensional shape[16,21,22]; the latter emphasises the significance of structure in drug discovery. However, large differences in potency between similar compounds – known as activity cliffs – can occur. While the cause in many cases remains unclear, activity cliffs can arise from steric clashes in the binding site, or the loss of a crucial interaction with the protein: problems that are less likely to arise in structure-based models.

When the three-dimensional structure of the target and location of the binding site are known, structure-based models require that the binding pose of a compound is determined before its affinity can be calculated. The process of positioning a compound in a binding site, known as "docking", has been the subject of research for over thirty years[23], and there currently exist many algorithms that are fast enough to be used in virtual screens[24]. Docking programs seek to find the binding mode that minimises, or maximises, a predefined score. The mathematical models used to calculate this score are called scoring functions. Computationally efficient protocols typically treat the protein as a rigid body, so that the docking program searches over the orientation of the ligand in the binding site and its dihedral angles. The size of the search space rises exponentially with the number of freely rotatable bonds, so to speed up the optimisation, many docking programs are partially stochastic and produce many candidate binding modes. Even though docking programs can produce accurate predictions within the candidate solutions, scoring functions often struggle to distinguish the best pose[25,26].

In addition to binding mode prediction, scoring functions are also used to produce estimates of the binding affinity. Some methods use the same scoring function as for binding mode prediction, while others decouple binding and affinity prediction and rescore the complexes with a different model[27]. Despite large amounts of research, it

remains a challenge to predict the affinity of a compound even when an experimental protein structure with the bound ligand exists[28]. Errors are even greater when predicting affinities from docked poses[29]. Although virtual screening with docking and scoring can significantly enrich compound libraries, the performance of a given scoring function is highly dependent on the data set[25,26,30–33].

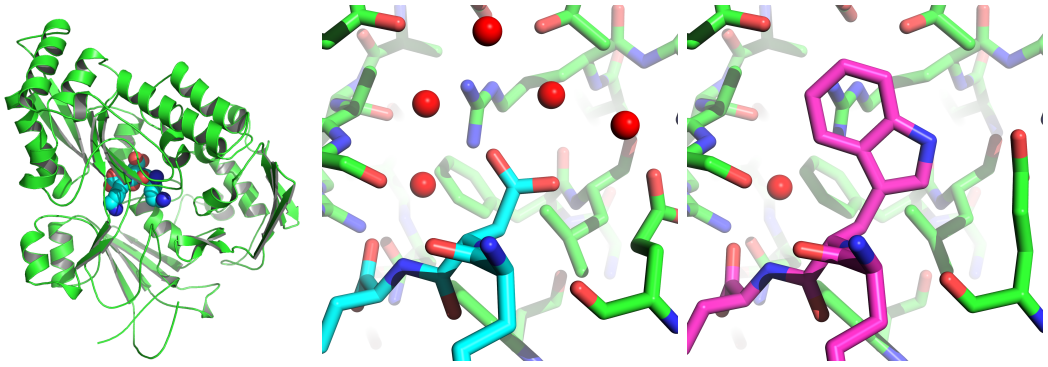
Scoring functions are the fastest structure-based method for estimating affinity, typically using a single docked pose and the rigid structure of the protein. In reality, protein-ligand complexes are dynamic, not static, and, as will be discussed in Chapter 2, the binding affinity is a function of the thermodynamic free energy difference between the bound and unbound states of the protein and ligand. While closely related scoring function methods have attempted to account for multiple ligand positions using the candidate solutions from the docked poses[34–36], they still do not account for the unbound state of the protein and ligand.

Atomistic simulations, either with molecular dynamics or Monte Carlo methods, can respectively generate a trajectory or an ensemble of structures of either the protein, ligand, or complex at comparatively great computational expense compared to docking programs. Methods such as the relaxed complex scheme approximate the binding free energy as the average interaction energy between the bound ligand and protein[37]. Alternatively, the energetics of a collection of structures from a simulation can be recalculated using Poisson-Boltzmann (PB) or generalised Born (GB) electrostatics, which straightforwardly estimate the binding enthalpies of the ligand and binding site, and can estimate the binding solvation entropy by further assuming the change is proportional to the surface area (SA) of the binding site and ligand. As a result, these techniques are known as PBSA and GBSA[38]. Computationally more expensive, so-called end point methods run atomistic simulations of the protein and ligand when they are bound *and* unbound. A popular end point formalism known as linear interaction energy[39] assumes

that the free energy is given by a linear response in the average difference in potential energy terms in both simulations.

It is important to stress that all of the aforementioned techniques require significant approximations or assumptions regarding the binding free energy. The theory of statistical mechanics – from which many of the above methods are derived, and which is outlined in Section 2 of Chapter 2 – also yields theoretically *exact* ways to calculate the affinity of a protein-ligand complex[40,41]. The increased rigour comes at the expense of significant computational resources, and often multiple simulations are required to calculate the affinity of a single complex. Potential of mean force (PMF) methods are an example of a rigorous technique; these calculate free energy profiles along pre-defined reaction coordinates. A set of rigorous techniques known as "alchemical" methods calculate the free energy difference between states, such as the bound and unbound states of a protein and ligand, using a series of non-physical intermediates, reminiscent of the long-sought- after elemental transmutation from which these techniques derive their name. Although binding sites that are buried within a protein make it difficult to define a reaction coordinate for a PMF calculation, alchemical techniques, such as thermodynamic integration, have no such difficulty as ligands can gradually be transformed in and out. An alternative alchemical method, known as free energy perturbation or exponential averaging, is most often applied to calculate the relative affinities of congeneric compounds[42].

In theory, rigorous binding free energy calculations are applicable to any protein-ligand complex. They are limited by forcefield approximations and inadequate sampling, problems that are also faced by the less rigorous simulation-based methods. However, they remain technically difficult and time consuming[41,43], and are more likely to be of use in the lead optimisation stage of drug discovery than in virtual screening, where the simpler ligand-based and structure-based scoring functions are widely used[6]. Interestingly, in the few cases where structure-based scoring functions and rigorous methods have been



(A) (B) (C)

FIGURE 1.3: (A) The bacterial peptide transporter OppA binds to a myriad of small peptides by adjusting the number of water molecules in its binding site[54]. (B) As the peptide (in blue) is too small, four water molecules (red spheres) are able to bridge the interaction to the protein (green sticks). (C) In the same binding site, all but one of the water molecules have been expelled to accommodate the larger peptide side chain (purple sticks).

compared, the improvement from the scoring function is slight[44,45], and tests of rigorous methods often report absolute errors of 1-3 kcal/mol[46], which are similar to the reported errors of scoring functions in prospective studies[47].

1.4 Water in binding thermodynamics

Water is a substance that is utterly fundamental to biomolecular function[48]. Among many of their structural roles, water molecules can stabilise proteins[49], lubricate molecular hinges[50] and control ion channel gating[51,52]. Pertinently, water's critical role in protein-ligand binding simultaneously informs lead optimisation strategies and affinity prediction methods. When binding in an aqueous environment, both the protein's binding site and the ligand partially desolvate to facilitate direct interactions between the pair. Where the protein and ligand do not directly interact, water molecules are commonly found to mediate interactions[53], allowing binding sites to accommodate a variety of ligand sizes and chemistry that otherwise would have been impossible (see Figure 1.3).

The inclusion of explicit water molecules in docking, scoring and rigorous free energy calculations is becoming ever more common[55–57]. Careful consideration of hydration sites has been shown to aid the predictability of 3D QSAR models,[58–61] ensure stable molecular dynamics simulations[49], and improve the accuracy of rigorous free energy calculations[62].

Despite their importance, locating where water molecules reside protein structures can itself be problematic. Discussed in more detail in Chapter 3, water molecules are often extracted from X-ray crystal structures, where they are added towards the end of the structure refinement procedure. There is no standardised method for including water molecules or validating their positions, such that different crystallographers can predict different water distributions from the same diffraction pattern[63]. The most rigorous way to computationally predict water molecule locations requires all-atom molecular simulations, whose unprocessed results are a continuum of positions. While a continuum may be the closest representation of water's "true" spatial extent, it is interesting to note that interpreting water's biomolecular role and incorporating water in quantitative models is most easily achieved when using a discretised set of explicit positions.

Ideally, as scoring functions predict affinity using only a single snapshot of a molecular complex, the objective of scoring function research is to elucidate any emergent laws that operate above the physical laws of atomic motion. The hydrophobic effect, the tendency for non-polar molecules to aggregate in water, is the best known emergent property of biomolecular association. The history of scientists' understanding of the hydrophobic effect is not only an exemplary story of how collaboration between experiment, theory and computational methods are needed to elucidate the thermodynamics of bimolecular binding, but it contains many lessons for advancing affinity models and structure-based drug design as well.

1.4.1 The hydrophobic effect

In 1945, Frank and Evans proposed the famous "iceberg" model of hydrophobicity[64]. By observing that the hydration of non-polar gases was accompanied by a decrease in entropy larger than that of ions, they reasoned that water forms ice-like cages around non-polar solutes in an effort to conserve hydrogen bonds. The increased ordering in water, they postulated, was responsible for the decrease in entropy. Further experiments with water at higher temperature showed that the decrease in entropy lessoned, apparently signifying the "melting" of clathrate cages[64]. With this model, non-polar aggregation could be understood by the liberation of caged water molecules upon association of two hydrophobes, such that the association would primarily be *entropically* driven. This picture was readily adopted by those seeking to understand why non-polar amino acids are typically found at the core of aqueous proteins[65,66]. In the 1980s, new experiments raised doubts about the classic iceberg model[67], such as the observation that protein-protein association could be *enthalpically* driven, even in cases where binding involved the burial of non-polar side chains[68]. Also, beginning in the late 1970s and confirmed later by molecular simulations, theoretical studies indicated that the increased ordering could *only* lead to the aggregation of non-polar solutes at high concentrations of solute[69–72].

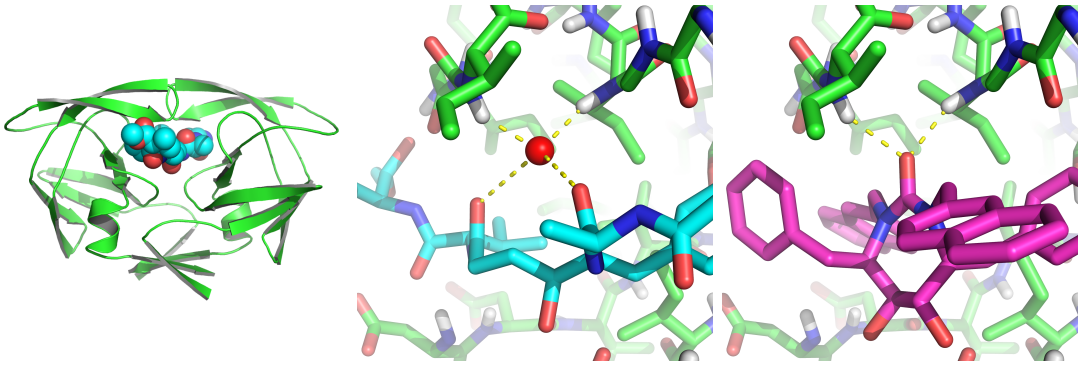
A significant breakthrough was the realisation that the driving forces of the hydrophobic effect are dependent on the size and curvature of the solute[69,71,73]. For instance, with large non-polar surfaces, it is geometrically impossible for water to surround the solute and maintain all its hydrogen bonds. Unable to satisfy their hydrogen bonds, water molecules adjacent to these solutes have been observed to be disordered and vapour-like[74,75]. The association of large non-polar surfaces frees these comparatively high-energy water molecules, so that the process is more favourable enthalpically than it is entropically. Atomistic simulations of spherical non-polar cavities in water also exhibit

the same disordered, low-density water as found at the interfaces of large solutes[76]. More recent thermodynamic calculations of spherical, non-polar ligand and binding site revealed that binding was enthalpically favourable and entropically unfavourable, due to the displacement of the disordered water layer at the binding site[77]. Experiments, theory and atomistic simulations have shown that, while the literal interpretation of clathrate cages around non-polar solutes is completely unrealistic, the orientations of water molecules around small non-polar solutes are indeed restricted compared to bulk water[48,67,78]. Presently, it is well established that both entropy and enthalpy play a role in driving non-polar solvents together in water[78–80], although the topic continues to remain controversial[81–83]. Apparently unaware of the developments over the last thirty years, biochemical and biophysical textbooks, with notable exceptions[80], continue to only espouse the rigid water cage model of hydrophobicity[84–87], no doubt to the detriment of the field.

1.4.2 Hydrophobicity in structure-based drug design

Due to the hydrophobic effect, the addition of hydrophobic groups to ligands tends to increase the binding affinity with the protein. However, as hydrophobicity drives *all* non-polar solutes together, a more hydrophobic compound is likely to be less specific and partition more readily into membranes[88]. As a result, more selective strategies – informed by the thermodynamics of the hydrophobic effect – have been developed to exploit water in structure-based drug design.

Just as with the classic iceberg model of Frank and Evans, early rationalisations of the role of water in drug affinity were informed by the entropic benefits of molecular association. In 1994, Dunitz reasoned that, due to the loss of translational and rotational freedom, it was entropically unfavourable – with a maximum cost of 7 kcal/mol – for water molecules to bind to macromolecules[89]. The implication was that medicinal chemists



(A) (B) (C)

FIGURE 1.4: (a) The protein HIV protease is required by the virus to create a complete virion. (b) An X-ray crystal structure of HIV protease with an inhibitor reveals that a water molecule is bonded to both the protein (green sticks) and the inhibitor (blue sticks). (c) As the first example of its kind, a new inhibitor was designed to displace the same water molecule to improve its binding affinity[90].

should design compounds that displace ordered water molecules from binding sites to exploit the favourable increase in entropy. Three months prior to Dunitz's hypothesis, Lam et. al. reported that a new type of potent inhibitor for HIV-1 protease had been successfully designed to displace and mimic the bonds of an ordered water molecule[90] (see Figure 1.4) . Since this landmark work, the targeted displacement of ordered water molecules by polar ligand groups has been a popular strategy in drug discovery[91–98]. However, attempts to displace water molecules may not be successful[95,98], and can lead to a decrease in affinity if the ligand is unable to fulfil the water's stabilising role[96,97].

With regard to the thermodynamics of targeted water displacement, some studies combining ITC and X-ray crystallography have confirmed the classical view of hydrophobicity, where the stepwise expansion of non-polar groups into non-polar groves – accompanied by the displacement of water – increases the binding entropy[97,99]. Yet there are a growing number of studies reporting that enthalpy, rather than entropy, is the dominant favourable factor that occurs when non-polar groups displace water molecules[100–103]. Of course, this should only be surprising if one maintains the entropy fixated explanation of the hydrophobic effect. As discussed, the thermodynamic driving forces are dependent on the size and geometry of the solutes, and freeing structurally disordered waters

can result in favourable enthalpic and unfavourable entropic changes. However, confirming whether the displacement of disordered water is responsible for the favourable enthalpy observed in ITC experiments is difficult, not least because many factors, such as changes in the protein and ligand flexibility, contribute to the thermodynamic signatures observed in ITC[104,105]. In addition, structurally disordered waters in protein binding sites can be invisible to X-ray crystallography, even at a high resolution, requiring specialised protocols to resolve them[106]. Fortunately, NMR and MD simulations *can* reveal whether a binding site or cavity is hydrated with disordered, low density water[107–109]. A well-studied example of the enthalpic hydrophobic effect in protein-ligand binding is the thermodynamic analysis conducted by Homans and co-workers on mouse urinary protein[110,111], which has a very hydrophobic binding site. ITC experiments on the binding of non-polar ligands revealed that during ligand binding, affinity was dominated by a favourable enthalpy change, which could not be explained by the tightening of the protein as observed by NMR. As MD simulations revealed the binding site was sub-optimally hydrated with disordered water molecules, they concluded that the displacement of the disordered water molecules was responsible for the favourable enthalpy.

1.4.3 Hydrophobic consequences for modelling

Understanding how water affects the affinity of a protein-ligand complex is a notoriously difficult problem. The hydrophobic contribution to the binding of a non-polar ligand to a protein can be enthalpic or entropic, depending on the geometry and size of the binding site and ligand. An intuitive assessment of the thermodynamics is complicated by the fact that binding sites contain both hydrophobic and hydrophilic regions. Although the degree of desolvation of the protein's binding site and ligand that occurs during binding undoubtedly contributes to the affinity of the complex, a recent analysis of thousands

of protein-ligand complexes revealed no statistically significant difference between the affinities of complexes with bridging water molecules and those without[112]. Further complicating a simple understanding of the thermodynamic contributions of water are cooperative effects[97], and the fact that the binding of a ligand may not directly displace water molecules from a binding site but rather perturb an adjacent water hydrogen bonding network[113].

The only way to predict the role of water seems to be with a computational approach, and great strides have been made in accounting for water's multifaceted effects in drug-protein binding, even in cases involving water network perturbation[114]. Rigorous free energy methods, such as thermodynamic integration, can calculate the binding free energies of water molecules in protein binding sites[115,116], and have been applied to predict the change in affinity when a ligand group is designed to displace an ordered water molecule[117,118]. Calculating the enthalpic and entropic contributions of water displacement with such rigorous methods is prohibitively time-consuming in a practical setting, requiring free energy calculations to be performed at different temperatures. From a single simulation, inhomogeneous solvation theory (IFST), as popularised by Lazeridis in the late 1990s[119,120], can calculate the binding entropy and enthalpy of ordered water molecules at binding sites. The entropic cost of binding a water molecule to a protein is estimated by truncating solvent-solute correlation functions, typically to the second order, in a manner similar to theories which appeared in the 1980s[121,122]. This method is the basis of the commercially available semi-empirical method WaterMap[123,124], which, among many applications, has been applied to understand the affinity and selectivity of kinase inhibitors[125] and PDZ domains[126]. IFST is typically applied to simulations of restrained biomolecules, and the impact on the thermodynamic estimates has been investigated by Huggins[127,128].

An early application of IFST to drug discovery related problems was conducted by Li

and Lazaridis on the water molecule displaced by Lam et. al. in HIV-1 protease[129] (shown in Figure 1.4). They found, that as expected, the water molecule was highly ordered compared to bulk water. The decrease in entropy for transferring the water molecule from bulk water was calculated to be 9.8 kcal/mol, exceeding the maximum of 7 kcal/mol initially estimated by Dunitz. The calculated solvation free energy of the bound water molecule was -15.2 kcal/mol, in contrast to -9.1 kcal/mol as calculated later by Hamelberg and McCammon with thermodynamic integration[115]. Calculations by WaterMap have confirmed that ordered water molecules in binding sites have a lower entropy with respect to the bulk solvent[123]. Recent applications have also revealed that ordered water molecules can have a lower enthalpy than bulk water, which were used to explain the favourable enthalpic component observed in the targeted water displacement in carbonic anhydrase II[100]. The water excluded regions within non-polar pockets, discussed above, fall outside the framework of IFST, prompting a recent extension of WaterMap[130].

1.5 Objectives for this thesis

This thesis concerns the accurate prediction of protein-drug affinities: one of the "holy-grails" of computational chemistry[7,18,131]. Of particular interest will be methods that can be used in virtual screens.

Two significant trends can exploited to develop more accurate, fast affinity prediction methods: the continual exponential rise in computational power as described by Moore's Law[132], and the ever-increasing availability of structural and thermodynamic binding data of protein-ligand complexes[133–137]. While Moore's Law suggests that methods which utilise currently time-consuming molecular simulations can one day be incorporated into virtual screenings, structure-based scoring functions appear to have the most

to gain from both trends, as faster processors facilitate the screening of ever larger libraries of compounds, and the large amounts of experimental data can be used to discover and exploit statistical and emergent properties to calculate affinities. This thesis, therefore, approaches the development of affinity estimation from an empirical perspective.

The hydrophobic effect is arguably the best studied emergent effect in biomolecular binding. As computational methods have proven essential in elucidating the thermodynamic contributions of water molecules in binding, a reasonable hypothesis is that a more detailed treatment of water in structure-based scoring functions can improve their accuracy over current scoring functions, which either ignore water or use simple implicit solvent models. Methods to rapidly determine water molecule locations and to predict their roles in protein-ligand binding will be pursued in order to contribute to the main goal of water-based scoring function.

A recurring theme in this thesis concerns the role of uncertainty in structure-based drug design and affinity prediction. As such, the central underlying assumptions of thesis – that increased amounts of experimental and structural data and improved water modelling can improve scoring functions – will be investigated. It is hoped that the successes and failures of this thesis will yield lessons and future directions toward the dream of rapid affinity prediction.

Chapter 2

Background theory

2.1 Binding affinity

Consider the simple bimolecular reaction $L + M \rightleftharpoons ML$, where L and M are the ligand and macromolecule respectively, and ML is the complex. This reaction is reversible, so that if the reactants were well mixed in a beaker of solvent, we may imagine the complex constantly assembling and disassembling. The relative concentrations of the free ligand and macromolecule to the concentration of the complex indicates the binding strength, and the affinity is quantified by the dissociation constant

$$K_d = \frac{[M][L]}{[ML]}, \tag{2.1}$$

where square brackets denote molar concentrations. For a high affinity, there would be a large concentration of the complex $[ML]$, compared to smaller amounts of the free macromolecule $[M]$ and ligand $[L]$; the smaller K_d is, the stronger the interaction. The inverse of K_d, known as the association, or equilibrium constant, is also a common

measure of affinity. However, K_d benefits from a simple interpretation: it is equal to the concentration of free ligand at which half of the macromolecule is bound. The speed that the reaction $L + M \rightleftharpoons ML$ occurs in either direction is hidden within K_d. Denoting the rate at which the complex forms as k_{on} and dissociates as k_{off},

$$K_d = \frac{k_{\text{off}}}{k_{\text{on}}}, \tag{2.2}$$

Thus, a ligand and a protein which bind with a high affinity may have a high association rate, or a low dissociation rate, or both.

As discussed in Chapter 1, the accurate prediction of K_d remains a challenging problem, the resolution of which would be of great benefit to drug discovery. Many of the most successful prediction methods have followed from a deeper understanding of the physical basis of binding affinity, which is the subject of the next section.

2.2 Thermodynamics and statistical mechanics

We can begin to understand the physical basis of binding affinity by looking at proteins and ligands at the microscopic scale. In the highest detail, the electrons, protons and neutrons of biomolecules are described by the probabilistic quantum theory of matter, whose dynamics are described by Schrödinger's or Heisenberg's equations of motion [138]. At a coarser level of resolution, quantum uncertainty relents, and the kinetics of atoms appears deterministic. At this scale, classical mechanics reigns and the atomic constituents of proteins and ligands are well approximated by Newton's equations of motion. To link this microscopic view of nature to the macroscopic world of lab benches and experimentally measured drug affinity requires classical determinism to be supplanted

by yet another probabilistic theory, where our uncertainty arises from the unimaginable – but countable – number of atoms and molecules that make up a living cell or a beaker of water.

In thermodynamics, systems are completely characterised by only a handful of variables, such as temperature and average total energy. A beaker of water left in a closed room would eventually equilibrate to the room's same temperature and pressure. If evaporation is slow, the number of water molecules it contains will be roughly constant. In this example, the thermodynamic variables are the temperature, denoted T, pressure P, and number of particles N. Knowing only so few variables means that we are ignorant of the positions, momenta and internal degrees of freedom of the N water molecules in the beaker.

The positions and momenta of the N particles are denoted by the coordinate vectors $\mathbf{r}^N$ and $\mathbf{p}^N$ respectively. The hyper-dimensional coordinate system spanned by $\mathbf{r}^N$ and $\mathbf{p}^N$ is known as the phase-space, and each point in phase-space represents a particular microscopic configuration of the system. The internal degrees of freedom possessed by a molecule, such as bond vibrations and rotations, also contribute to the dimensions of phase-space, but are omitted in this discussion for brevity. The traditional approach to gaining a statistical understanding of a system – covered more rigorously in textbooks such as McQuarrie[139] – imagines infinitely many copies of the same system, each at different snap-shots in time. The collection of systems is known as an *ensemble*, and different ensembles can be constructed for different sets of macroscopic constraints. By assuming that microscopic states of equal energy are equally probable, and that time averages are equal to averages over the ensemble, a probability distribution over phase space, known as the Boltzmann distribution, can be established. A fixed T, P and N corresponds to the isothermal-isobaric ensemble, and it can be shown that the probability over phase-space is given by the following Boltzmann distribution

$$p_B(\mathbf{r}^N, \mathbf{p}^N) = \frac{1}{\Delta(T,P,N)} \exp\left(-\frac{E(\mathbf{r}^N, \mathbf{p}^N)}{k_B T} - \frac{PV}{k_B T}\right), \tag{2.3}$$

where $E(\mathbf{r}^N, \mathbf{p}^N)$ is the energy at a particular point in phase-space, V is the volume, k_B is Boltzmann's constant and $\Delta(T,P,N)$ is the normalisation factor, known as the partition function. It is given by

$$\Delta(T,P,N) = C(N) \int_0^\infty \exp\left(\frac{-PV}{k_B T}\right) \int_V \int_{-\infty}^{+\infty} \exp\left(\frac{-E(\mathbf{r}^N, \mathbf{p}^N)}{k_B T}\right) \mathrm{d}\mathbf{r}^N \, \mathrm{d}\mathbf{p}^N \, \mathrm{d}V \tag{2.4}$$

where $C(N)$ ensures the non-dimensionality of $\Delta(T,P,N)$, consistency with quantum mechanics, and accounts for whether the particles are distinguishable or indistinguishable. An alternative way to approach statistical mechanics – discovered by Jaynes [140,141] and covered in textbooks such as Dill and Bromberg [80] – does away with the construction of thermodynamic ensembles, and derives the above statistics from a purely Bayesian perspective. In Jaynes' framework, maximising the *uncertainty* of the phase-space distribution, subject to the macroscopic constraints, yields the Boltzmann distribution.

From $\Delta(T,P,N)$, all of the macroscopic observables of the system can be derived. The above, therefore, provides the bridge between the microscopic and the macroscopic world. All macroscopic observables are given by, or are related to, expectation values over relevant Boltzmann distributions. The macroscopic value of a microscopic variable $A(\mathbf{r}^N, \mathbf{p}^N)$ is given by the average over all phase space

$$\langle A \rangle = \frac{1}{\Delta(T,P,N)} \int_0^\infty \exp\left(\frac{-PV}{k_B T}\right) \int_V \int_{-\infty}^{+\infty} A(\mathbf{r}^N, \mathbf{p}^N) \exp\left(\frac{-E(\mathbf{r}^N, \mathbf{p}^N)}{k_B T}\right) \mathrm{d}\mathbf{r}^N \, \mathrm{d}\mathbf{p}^N \, \mathrm{d}V, \tag{2.5}$$

where, in general, angular brackets denote ensemble averages. The binding affinity between a drug and a protein is also a macroscopic observable. To relate protein-ligand binding affinity to statistical mechanics, it informative to consider the Gibbs free energy of the system

$$G = -k_B T \ln \Delta(T,P,N), \tag{2.6}$$

where the dependency of G on T, P, and N is made implicit for notational simplicity. The Gibbs free energy equals the maximum amount of work that can be extracted from a system for a fixed P and T. The free energy can be expanded

$$G = H + TS, \tag{2.7}$$

where H and S are the enthalpy and entropy, respectively, of the system. The enthalpy

$$H = \langle E \rangle + P\langle V \rangle, \tag{2.8}$$

quantifies the average energetic contributions of the system and the maximum amount of heat that can be extracted from it. The entropy

$$S = -k_B \int_V \int_{-\infty}^{+\infty} p_B(\mathbf{r}^N, \mathbf{p}^N) \ln p_B(\mathbf{r}^N, \mathbf{p}^N) \, \mathrm{d}\mathbf{r}^N \, \mathrm{d}\mathbf{p}^N, \tag{2.9}$$

can be understood as the degree of microscopic uncertainty, as the log of the average phase-space volume sampled by a system is proportional to S. The greater the entropy, the worse our ability to specify the microscopic configuration of the system. The entropy is maximised for uniform distributions, and as Equation 2.3 shows, the larger the temperature becomes, the more $p_B(\mathbf{r}^N, \mathbf{p}^N)$ tends to a uniform distribution, and the more disordered the system is said to be.

Consider again the binding reaction $L + M \rightleftharpoons ML$, between a ligand L and macromolecule M. This reaction implies two distinct macrostates: the unbound state, where the ligand and macromolecule are separate, and the bound state, where both molecules have formed a complex. One can associate a free energy to both of these states, the unbound free energy G_u and the bound free energy G_b. The difference between the bound and unbound free energies

$$\Delta G = G_b - G_u, \tag{2.10}$$

is known as the binding free energy. The sign and magnitude of ΔG determines the likelihood of finding the macromolecule and ligand in the bound or unbound states. If ΔG is negative, the binding reaction can occur without any additional input of energy, and it is more probable to find the molecules in the bound state than the unbound

state. A seminal result in statistical mechanics is the following relationship between the dissociation constant and the binding free energy:

$$\Delta G = -k_B T \ln \frac{K_d}{K_{\text{ref}}}, \tag{2.11}$$

where K_{ref} is the reference concentration, typically set equal to 1 mol/litre. For a binding reaction occurring at the same temperature, Equation 2.7 implies that $\Delta G = \Delta H + T\Delta S$, so that the change in free energy can be understood in terms of the differences in enthalpy and entropy. As discussed in Chapter 1, Section 1.4.1, the consideration of the entropy and enthalpy changes that occur when two non-polar solutes aggregate in water is crucial to understanding the driving forces behind the hydrophobic effect.

Together, Equation 2.4, Equation 2.11 imply that if one were to calculate the energies for all of phase-space in the bound and unbound macrostates, one could predict the binding affinity of a protein-ligand complex. Molecular simulations, such as molecular dynamics and Monte Carlo methods, use forcefields to estimate the energies at particular points in phase-space. However, owing to the sheer enormity of phase-space, and the vanishingly small probability that energetically unfavourable regions will ever be sampled[142], brute force calculations of G_b and G_u are out of the question for all but the simplest toy problems. Instead, dedicated binding free energy techniques, such as free energy perturbation and thermodynamic integration offer more practical ways to calculate ΔG. These methods and others are covered in a number of reviews[40,41,143].

Many binding free energy methods have been inspired by expanding the binding free energy into independent components. When a ligand and macromolecule bind in an aqueous environment, a possible expansion is

$$\Delta G = \Delta G_{\text{solvent}} + \Delta G_{\text{protein-ligand}}, \tag{2.12}$$

where $\Delta G_{\text{solvent}}$ is the change in the water's free energy, which is contributed to by the partial desolvation of the ligand and binding site, and $\Delta G_{\text{protein-ligand}}$ is the change in free energy arising from interactions between the protein and ligand. Described in Chapter 1, Section 1.3, PBSA and GBSA techniques approximate $\Delta G_{\text{solvation}}$ by modelling water implicitly. Structure-based scoring functions are far more approximate techniques that, in the same spirit as statistical mechanics, base their binding free energy predictions on the microscopic detail of protein-ligand interactions. The majority of these models expand $\Delta G_{\text{protein-ligand}}$ with interaction energies or heuristic potentials, which can be evaluated as quickly as possible. Regression is often used to train these simpler models on the structures of protein-ligand complexes. The theory underlying regression is covered in the following section.

2.3 Statistical learning theory and regression

Statistical mechanics is a triumph of deductive reasoning, as the physics at the microscopic level can be used to predict binding affinity, a macroscopic phenomenon. In contrast, one can approach affinity prediction inductively, and investigate which structural features of protein-ligand complexes are indicative of high or low affinity.

Inductive learning requires prior data, known as a training set, on which a model can be based. In structure-based scoring functions, training sets are comprised of protein-ligand complexes with experimentally measured binding affinities and resolved X-ray crystal structures. Statistical learning theory (see Hastie et. al.[144]) provides a framework under

which functional relationships between features and affinities can be reliably inferred. The process is known as regression, and models obtained in this way inherit the assumptions that are implicit in statistical learning theory. In regression, features or descriptors – such as shape complementarity in scoring functions – are known a explanatory variables, which we denote as X, and the predicted variable – like binding free energy – is called the response variable, denoted Y. Following the common convention in statistics, we will denote random variables with capital letters and individual realisations of those variables in the corresponding lower case. Lower case letters will also be used within functions.

A criterion known as a cost or loss function is used to judge the success of a model. A very popular criterion for continuous response variables, like binding free energy, is the mean squared error (MSE) between the predictions of the model and the measured values. To complicate matters, the best model is unlikely to be the one with the lowest error on the training set. An arbitrarily complex function that passes through all the data points in a training set will have zero error, but may be disastrously wrong in predictive setting. The greater the non-linearity of the fitted function and the more free parameters it has, the more likely it is the model has over-fit the training data. The goal of regression, therefore, is to find the function that *consistently* has the lowest error on hitherto unseen data. Statistical learning theory encapsulates this concept of reproducibility by assuming there is a probabilistic process underlying the data. A training set, for instance, is modelled as multiple samples from a joint probability distribution function (PDF) over X and Y, which we denote as $p(x, y)$. Equally, a test set is modelled as another set of samples from $p(x, y)$. Thus, by training a model to have a low error in a predictive setting, we are forced to assume that any future data must be generated by the same joint PDF as the training set. By using regression to infer the functional relationship between X and Y, we are in actuality inferring something about the statistical relationship between the

two variables, which is encoded in $p(x, y)$. The expected MSE of an estimated function, denoted $\hat{f}(x)$, is given by

$$\text{MSE} = \iint_{XY} p(x, y) \left(y - \hat{f}(x)\right)^2 \, \mathrm{d}x \, \mathrm{d}y. \tag{2.13}$$

It is this quantity – not the average error on the training set – that one should seek to minimise when performing regression. A classic result in statistical learning theory is that the function that minimises Equation 2.13, denoted $f(x)$, is given by

$$\begin{aligned} f(x) &= \int_{Y} p(y|x) \, y \, \mathrm{d}y \\ &= E[Y|X = x], \end{aligned} \tag{2.14}$$

where $p(y|x) = p(x, y)/p(x)$ and $E[Y|X = x]$ denotes the conditional expectation value of Y given x. Thus, the optimal model is the conditional mean of the response variable for given particular explanatory variables. It seems then, that the regression problem is solved: to predict the affinity of a protein-ligand complex, for example, we need to provide the average value of Y for the particular structural features of the complex. However, in most real problems, such as with scoring functions, the distribution $p(x, y)$, which underlies the data, is unknown, so that such averages cannot be carried out. A great deal of statistical learning theory concerns finding reliable ways to estimate equations 2.13 and 2.14.

While the expected MSE cannot be evaluated directly, a number of regression strategies are informed by the bias-variance decomposition of the MSE. To proceed, we assume –

as Equation 2.14 implies – that $y = f(x) + \epsilon$, where ϵ is random noise with mean equal to zero and a standard deviation equal to σ_Y. For a fixed regression method, completely re-sampling the training set will result in a slightly different function estimate, $\hat{f}(x)$, than before, and we denote the average estimate due to complete re-sampling as $E[\hat{f}(x)]$. For a particular test set sample (x_t, y_t), the MSE can be expanded in terms of $\hat{f}(x_t)$ and $E[\hat{f}(x_t)]$to give

$$\text{MSE} = \sigma_Y^2 + \overbrace{\left(f(x_t) - E[\hat{f}(x_t)]\right)^2}^{\text{Bias}^2} + \overbrace{E\left[\left(\hat{f}(x_t) - E[\hat{f}(x_t)]\right)^2\right]}^{\text{Variance}}, \tag{2.15}$$

where the first term is the intrinsic noise in Y, which is irreducible; the second term is the squared bias, which quantifies how much the average of our estimate $\hat{f}(x)$ differs from the optimal function; the third quantity is the variance of the estimate, which tells us by how much our estimate is expected to change if we re-sampled the training set.

For every method of regression, there is a trade-off between bias and variance. A linear model will have a very high bias if the true relationship between the explanatory and response variables is not linear. However, linear models have a lower variance than non-linear models precisely because the model's functional form is significantly constrained. Highly non-linear models, such as neural-networks, have a very low bias because they make few assumptions regarding the functional relationship between the variables. This higher complexity comes at the cost of larger variance. Such models are sensitive to the composition of the training and test sets, and re-fitting a non-linear model again to a re-sampled training set can result in a very different function.

Bootstrap aggregating (commonly known as bagging) is a common technique that is applied to reduce the variance of complex models, and is often used in conjunction with regression trees. In bagging, a large number of "new" training sets are obtained

by sampling uniformly with replacement from the original training set to approximate complete re-sampling. A model is fit to each sample, and future predictions are given by the average over the fits. With reference to Equation 2.15, the average over all the fitted models is closer to the $E[\hat{f}(x)]$ than any single model, so that the "bagged" models have a lower variance.

It is important to highlight that the definition of an optimal model is dependent on the choice of loss function. Here, we discuss the MSE due to its popularity and analytical simplicity, but one could equally choose another loss function, such as the mean absolute error (MAE). It can be shown that the function that minimises the MAE is given by the conditional *median* of the response variable, as opposed to the conditional *mean* given in Equation 2.14. Depending on what type of distribution $p(x, y)$ is, these quantities may be very different.

In addition to the error of a model, the degree of correlation between a model's predictions and the experimentally measured outcomes is another easily interpretable quantification of predictive success. Pearson's linear correlation coefficient,

$$R = \frac{\text{cov}[X, Y]}{\sigma_X \sigma_Y} \tag{2.16}$$

is one such measure, where $\text{cov}[X, Y]$ is the covariance between X and Y, and σ_X and σ_Y are the standard deviations of X and Y respectively. It has a maximum value of 1, indicating a perfect positive linear correlation, a minimum of -1 for a perfect negative linear correlation, and equals 0 in the case of no linear correlation. The square of R is equal to the amount of linear variance in Y explained by X. In drug-discovery, however, one of the most important measures of a scoring function's success is its ability to correctly rank order compounds by affinity, as in a virtual screen, only the top ranked

compound are typically selected for experimental testing. Spearman's rank coefficient, often denoted as ρ, measures how well the relationship between two variables can be measured as a monotonic function. It is given by Pearson's correlation coefficient of the rank ordered predictions and experimental values. A value of 1 means that all the predictions have the same rank ordering as the measured values. In general, Pearson's and Spearman's coefficients are not the same. For instance, a monotonic relationship between the predictions and measured values with a high Spearman rank coefficient need not be linear, and so may have a low Pearson correlation coefficient.

2.4 Information theory and regression

Originally conceived by Claude Shannon as a general theory for communication, information theory has wide ranging applications, including statistical mechanics and regression, the latter which we primarily discuss in this section. At its heart, the theory provides a measure for the information content of a given system. Information theory is well covered by Cover and Thomas[145].

To gain an intuitive understanding of information as a quantity, consider the question "What is the information content of this thesis?". Such a question concerns only the amount of information, not its quality or importance. The sum of all the text this thesis contains would be a very poor measure, because any repeated words or phrases would not be informative, having already been covered by preceding text. Thus, the text that appears with a higher frequency should, on average, contribute to the thesis' information content less than the text that occurs infrequently. Viewing the thesis as a probabilistic source of letters, or words, or phrases, denoted X, that are produced with probability $p(x)$, we would expect that a particular X that occurs with a low probability provides

more information on average than a realisation of X that occurs with a high probability. In his landmark paper[146], Shannon reasoned that for discrete probability distributions,

$$H(X) = -\sum_{X} p(x) \log p(x), \tag{2.17}$$

is the amount of information provided by a source whose output is governed by $p(x)$. This quantity, known as the Shannon, or information entropy, is always non-negative. As information theory was originally applied to electrical signals, the logarithm is typically taken to the base two, so that entropy is measured in *bits*. Equation 2.17 shows that the entropy is the average value of $\log \frac{1}{p(x)}$, implying that less probable outcomes are weighted higher than more likely outcomes, as discussed. Entropy is a property of the probability distribution only, which may be different for distinct message sources. For continuous random variables, which we will henceforth consider, the differential or continuous entropy is defined as

$$h(X) = -\int_{X} p(x) \ln p(x) \, \mathrm{d}x, \tag{2.18}$$

where in contrast to the discrete case the natural logarithm is commonly used, and $p(x)$ now denotes a PDF. In addition to quantifying the amount of information contained in a message, the entropy characterises the degree of spread in a distribution, or equivalently, the "uncertainty" in a random variable. Accordingly, $h(X)$ is at a maximum for uniform distributions. Shannon entropy derives its name from Gibbs entropy in statistical mechanics. Comparing equations 2.9 and 2.18, one can see that the Gibbs entropy can be considered as the information content of a thermodynamic system.

In regression, one is concerned with two types of random variables, the explanatory variables and the response variable. The goal of regression is to find the function that maps the explanatory variables, denoted X, to the response variable, denoted Y, with the lowest expected error. While regression theory states that the function that yields the lowest MSE is given by Equation 2.14, it does not state how one chooses the explanatory variables in the first place. A possible way is via the mutual information, $I(X;Y)$, which quantifies the amount of information shared by X and Y. It is given by

$$I(X;Y) = \iint_{XY} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \, \mathrm{d}x \, \mathrm{d}y, \tag{2.19}$$

and is a measure of statistical dependence between X and Y. It is positive unless X and Y are independent, in which case $p(x,y) = p(x)p(y)$, so that $I(X;Y) = 0$. To understand $I(X;Y)$ further, one can expand

$$I(X;Y) = h(Y) - h(Y|X), \tag{2.20}$$

where

$$h(Y|X) = -\iint_{XY} p(x,y) \ln p(y|x) \, \mathrm{d}x \, \mathrm{d}y, \tag{2.21}$$

is the conditional entropy, which quantifies the uncertainty in Y given X, and tells us how much information Y conveys if X is already known. A perfect set of descriptors could be used to determine Y unambiguously, in which case $h(Y|X) = 0$ so that $I(X;Y) = h(Y)$,

meaning that the best set of explanatory variables contain all the information in Y. For a fixed $h(Y)$, the minimum of $h(Y|X)$ occurs at the maximum of $I(X;Y)$. Concerning the statistical relationship between the variables, $h(Y|X)$ places an lower-bound on the MSE of any function obtained by regression. When X and Y are non-dimentional,

$$\text{MSE} \geq \frac{1}{2\pi e} \exp\big(2h(Y|X)\big), \tag{2.22}$$

with equality if and only if $p(y|x)$ is a normal distribution. Thus, there is an important link between the information variables share with each other and the optimal performance of empirical models.

The Kullback-Leibler divergence, or relative entropy, between two distributions is an information theoretic term that has also impacted regression analysis. For two PDFs over X, $p(x)$ and $q(x)$ it is defined as

$$D\big(p(x)||q(x)\big) = \int_X p(x) \ln \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{2.23}$$

The relative entropy measures the dissimilarity between two distributions, it is zero when $p(x) = q(x)$ and positive otherwise. Considering $D\big(p(x)||q(x)\big)$ as a type of distance between $p(x)$ and $q(x)$ is illustrative, although not strictly true, as it does not obey the triangle inequality, and is generally asymmetric with respect to exchanging $p(x)$ and $q(x)$. There are numerous, equally valid interpretations of the relative entropy. In the context of message processing and compression, the relative entropy is equal to the extra number of bits that are needed to compress a message governed by $p(x)$ when it is thought the distribution is $q(x)$[145]. In statistics, it is the average log-likelihood ratio

between two hypotheses, as well as the divergence between a model $q(x)$ and the true probabilistic process $p(x)$. Consideration of the latter led to the development of Akaike's information criterion (AIC) – covered in detail by Burnham and Anderson[147] – which is used to select between models of differing complexity if only a small training set is available. The AIC is an approximation of the relative entropy between reality and a model. For a fitted model,

$$\mathrm{AIC} = 2d - 2\ln(L) \tag{2.24}$$

where d is the number of free parameters in the model and L is the likelihood function of the fitted model. As discussed in the previous section, the error a fitted model has on a training set error typically underestimates the true error. The AIC corrects for the over-fitting of more complex models by penalising for the number of free parameters they contain. Instead of choosing the model the with the lowest training set error, one seeks to find the model with the lowest AIC.

Chapter 3

Rapid and accurate placement of water in protein binding sites

3.1 Introduction

The location of water molecules in binding sites and protein-ligand complexes has important implications in structure-based drug design. As discussed in Chapter 1, Section 1.4, water molecules can be targeted for displacement to improve affinity, increase specificity, and inspire new molecular scaffolds[90,125,148].

The positions of water molecules are typically taken from X-ray crystal structures and may be validated by comparison with related structures[149]. Nevertheless, there are inherent problems with identifying hydration sites in crystallography. Water molecules can be artifactual, may be too mobile to identify or not observed at all because of low resolution[63,107,109,150–152]. Structures resolved by NMR or predicted using homology modeling may be completely devoid of water molecules. Hence, it is necessary to be able to accurately predict water locations within binding sites.

Water sites can be predicted by running molecular dynamics or Monte Carlo simulations with an explicit water model and taking the peaks in water density or averaging over water molecule locations[153,154]. This has the benefit of including entropic effects in the prediction but can be very time consuming to run, especially with buried cavities due to the long time it takes for water to permeate within proteins. Grand canonical Monte Carlo methods, as utilised by the water thermodynamics package WaterMap[130], can significantly reduce the length of the simulation[155], although can still be computationally demanding. The grid-based Monte Carlo method JAWS attempts to strike a balance between rapid solvation techniques and full molecular simulations that explicitly treat entropic effects[156]. It has the added advantage of producing an estimate of the free energy of displacing the water molecule into bulk solvent, although the value may not be well converged[41]. Integral theory approaches have also been developed to predict the structure of water in proteins[157–159]. Notably, the 3D reference interaction site model (3D-RISM), has been adapted to predict explicit water locations as well as their binding free energies and entropies[160], taking around 11 minutes on an eight core processor for a typical protein-ligand complex[161]. While a promising development, this still remains too slow to be applied in a virtual screening context.

Fast solvation methods have also been pursued for a number of years. A popular method is GRID, which calculates the interaction energy of a chemical probe around a protein[162]. The water probe is able to make up to 4 hydrogen bonds with the protein[163]. A novel mean field method has been reported by Setny and Zacharias that places potential water sites on a lattice and iteratively solves the solvent distribution using a semi-heuristic cellular automata approach[164]. The fact that water sites form distinctive distributions around amino acids[165] has been exploited by a number of knowledge-based methods[166–170]. An early example called AQUARIUS predicted solvent sites within a protein by mapping each amino acid to a data set of crystal structures[166]. SuperStar is

another knowledge-based method that combines structural data from the Protein Data Bank[171] (PDB) and the Cambridge Structural Database (CSD)[172] to predict chemical propensity maps within protein cavities[167]. Also using water distributions from the CSD, AcquaAlta predicted 66% of crystallographic water molecules in a test set of fourteen tripeptide ligands bound to OppA[169]. Favourable water bonding geometries have also been utilised to predict water sites[164,173]. By systematically analysing hydrogen bonding angles, Huggins and Tidor correctly predicted 44% of the bridging water molecules in a test set of twenty-one crystal structures of tri-peptides bound to OppA[174].

Clearly, there exist a great variety of computationally inexpensive ways to predict the location of water molecules in proteins. Unfortunately, many of the prediction methods are limited to predicting water molecules coordinated by polar groups[168,169,174], water distributions and not explicit locations[167], either bridging water molecules only[169,174] or apo hydration sites only[164,166], and can be monetarily costly[123,162]. With regards to the validation procedures of each method, comparisons to random placements are seldom reported[166,173], and false positive rates are rarely considered. Thus, if one wishes to predict the location of water in a binding site, not only are there a baffling array of possible methods to choose from and learn, but the accuracy of each one can be unclear.

Given the familiarity medicinal and computational chemists have with protein-ligand docking software, it is surprising that they have not widely been used to predict hydration sites[175]. Designed to solve multidimensional optimization problems, docking programs should be well adapted at balancing the various energetic needs of a water molecule. Using a method we call WaterDock, in this chapter we show that the AutoDock Vina tool[176] can be used to predict the location of ordered water molecules in ligand binding sites to a high degree of accuracy. Crucially, a WaterDock prediction only takes a matter of seconds to produce, utilises freely available software and is simple to implement. Its speed and accuracy mean it has the potential for use in virtual screening.

3.2 Method

3.2.1 AutoDock Vina as a water placement tool

Autodock is the most prominent brand of freely available docking software, and is widely used in the computational chemistry community. A water placement tool based on AutoDock would therefore be widely accessible, with many users being already familiar with its basic implementation.

AutoDock Vina (henceforth referred to as Vina) is the most recent version of the AutoDock series, being immediately preceeded by AutoDock 4. Docking with Vina is roughly two orders of magnitude faster than docking with AutoDock 4; in Vina's early validation, it was found that docking a compound with Vina took on average 8 minutes to complete, while AutoDock 4 took around 520 minutes[176]. Speed can be further increased in Vina as, unlike AutoDock 4, it allows for multithreading. The significant speed difference between the two was observed in our initial testing of docking with water. Our early comparisons also found that Vina was more able to predict a greater diversity of water sites in a single docking run compared to AutoDock 4. As such, Vina was taken forward for further investigations.

Vina employs a stochastic global optimisation procedure to locate the lowest energy binding poses, where the energy is calculated by Vina's scoring function. Vina's scoring function is composed of Gaussian shaped energy wells, steric repulsion, as well as hydrophobic and hydrogen bonding terms. The hydrogen bonding term accounts only for the proximity of hydrogen donor-acceptor pairs, and not the explicit location of the hydrogen atoms. This helps the energy landscape be smoother than that of AutoDock 4's forcefield-based scoring function, which includes among other terms Coulomb and Lennard-Jones potentials, and a directional hydrogen bonding term. While AutoDock 4 uses a Lamarckian genetic algorithm to find the lowest energy binding mode, Vina's

less coarse energy landscape aids the gradient-based optimisation stage of docking. The choice of Gaussian shaped energy wells in Vina reflects its shape complementarity-based approach to docking.

When docking a ligand, Vina begins by randomly placing the ligand within the specified search space. Then, successive steps of random perturbations – which are accepted or rejected according to the Metropolis criterion – are followed by local gradient-based optimisation via a quasi-Newtonian method. The total number of steps is automatically chosen by Vina, and is dependent on the degrees of freedom of the ligand, while the number of starting configurations is set by the user with Vina's "exhaustiveness" parameter. During optimisation, Vina records low energy binding modes, which are later clustered and refined using an undocumented internal procedure. Vina is limited to producing a maximum of 20 binding mode predictions for a given run.

The application of Vina to water docking requires several important considerations. Firstly, Vina's scoring function has been designed and optimised on drug-like compounds, not for water molecules as intended here. Thus, a training set was assembled to determine the applicability of Vina's scoring function to water. Secondly, Vina's maximum limit of 20 binding modes and internal clustering procedure means that a single long run with a high exhaustiveness parameter is not equivalent to many independent docking runs with respect to the number and diversity of the pose predictions. Thus, a second data set was assembled to optimise the docking procedure in terms of number of repeats, exhaustiveness of the search and the post-processing method. This second data was created with an emphasis on determining placement accuracy in terms of false positive and true positive rates. Finally, an independent test set was used to assess the accuracy of the final docking procedure, which we refer to as WaterDock.

3.2.2 Applicability of Vina's scoring function

The first data set comprised of 15 high-resolution, pharmacologically relevant protein crystal structures and is shown in Table 3.1. As there can be some inconsistencies regarding crystallographically observed water molecules, it may be that Vina correctly predicts hydration sites that are not observed experimentally. For this reason, three proteins from Table 3.1 were chosen for molecular dynamics (MD) simulations. The minimum distances from predicted water molecules to an experimental or a MD water molecule were used to investigate the relationship between a prediction's error and its Vina score. In order to assess the magnitude of the errors, the minimum distances were compared to those from a random placement of water molecules. The energy cutoff was chosen as the Vina score which produced an error distribution that was indistinguishable from a random placement.

Table 3.1 includes apo and holo crystal structures of some of the same proteins in order to test whether Vina can predict the location of bridging water molecules as well as water molecules in unliganded binding sites. The proteins were also selected to have a diverse number of water molecules in the binding site. For example, trypsin has only one water molecule bridging the interaction between the ligand (benzamidine) and the protein, whereas heat shock protein 90 has 9 bridging water molecules and 6 neighboring waters with its ligand, adenosine diphosphate (ADP). The unliganded structures of heat shock protein 90, penicillopepsin and PIM1 kinase were simulated using unrestrained MD for 10 ns. These proteins were selected as their binding sites vary in their hydrophobicity and are easily accessible to the bulk solvent. The MD simulation details are included in Appendix A.1.

Protein	PDB code	Resolution (Å)	Ligand
BRD4	2OSS	1.35	None
BRD4	3MXF	1.6	JQ1
Trypsin	1SOQ	1.0	None
Trypsin	1BTY	1.5	Benzamidine
HSP 90	1AH6*	1.8	None
HSP 90	1AM1	2	ADP
Penicillopepsin	3APP*	1.8	None
Penicillopepsin	1BXQ	1.4	PPi3
PIM1 kinase	1YWV*	2.0	None
PIM1 kinase	1XWS	1.8	BI1
PNPase	1V48	2.2	DFPP-G
GluA2	1FTM	1.7	AMPA
HIV-1 protease	1KZK	1.1	JE-2147

TABLE 3.1: The protein structures used to establish a cutoff score that indicates whether or not a prediction is better than random. *Structures selected for molecular dynamics simulations.

Table 3.1 includes apo and holo crystal structures of some of the same proteins in order to test whether Vina can predict the location of bridging water molecules as well as water molecules in unliganded binding sites. The proteins were also selected to have a diverse number of water molecules in the binding site. For example, trypsin has only one water molecule bridging the interaction between the ligand (benzamidine) and the protein, whereas heat shock protein 90 has 9 bridging water molecules and 6 neighbouring waters with its ligand, adenosine diphosphate (ADP). The unliganded structures of heat shock protein 90, penicillopepsin and PIM1 kinase were simulated using unrestrained MD for 10 ns. These proteins were selected as their binding sites vary in their hydrophobicity and are easily accessible to the bulk solvent. The MD simulation details are included in Appendix A.1. For each crystal structure or MD snapshot, Vina was used to dock

a single water molecule into the binding site and all the locations were recorded. To ensure a sufficient diversity of water sites, Vina was used twice on each structure and the maximum number of predicted binding modes was retained, producing 40 water site predictions for each binding site. Prior to docking, structures were stripped of all pre-existing water molecules and prepared using the software package AutoDockTools[177]. For holo-proteins, the search space was 15Å×15Å×15Å around the geometric centre of the ligand. Apo-proteins were structurally aligned to the corresponding holo structure and the ligand centre was again used to define the docking search space. This search volume is sufficient to encompass the binding sites of the proteins listed in Table 3.1.

As stated, Vina's predictions were compared to a random distribution of water molecules. Water molecules were placed at random within the sterically allowed volume of each docking search space. AutoGrid (part of the AutoDock 4 package)[178] was used to create oxygen affinity grid maps, and favourable points were selected at random on grid locations that had energies less than or equal to 0 kcal/mol. Five hundred randomly placed water molecules were selected for each protein structure.

3.2.3 Refinement of the water placement method

Repeated independent water molecule dockings creates many overlapping and similar water predictions even after low energy sites have been removed. A second data set was created in order to test the accuracy of different clustering methods and docking procedures. An accurate water placement method is one in which many experimental water positions are correctly identified (high true positive rate) with very few predictions that are not experimentally observed (low false positive rate). As discussed in Section 3.1, the validity of water molecules seen in X-ray crystal structures is often uncertain and many water molecules may be missing from the structure. This complicates the proper assessment of the sensitivity and specificity of a water placement method. To

circumvent these issues, the data set in Table 3.2 was assembled so that each structure had been determined to a high resolution more than once. Where possible, neutron diffraction data were included because of its ability to resolve water molecules that contain deuterium. Each protein in Table 3.2 was structurally aligned and consensus water molecules were determined. A consensus water molecule was defined as one that was within 1 Å of another water molecule seen in at least one other structure. These water molecules were used to estimate the true positive rate of WaterDock. The binding site water molecules that were seen in only one structure were recorded in order to estimate the false positive rate of WaterDock. By validating WaterDock in this way, WaterDock's true positive rate was assessed using only trustworthy water sites, while its false positive rate was assessed using all water sites, for which there is at least some evidence for. Note that because of the difficulty in experimentally resolving some water molecules, the false positive rate is likely to be an upper estimate.

Protein	PDB codes	Resolution (Å)	Ligand
HIV-1 protease	3FX5, 1HPX, 2ZYE*	0.9, 2, 1.9	KNI-272
Ribonuclease A	1KF5, 1FS3, 5RSE*	1.2, 1.4, 2	None
GluA2	1FTM†, 1MY2†	1.7, 1.8	AMPA
Trypsin	1S0Q, 1UTQ, 1TPO*	1.0, 1.2, 1.7,	None
Concanavalin A	1NLS, 1GKB, 1JBC, 1QNY*, 1K3L	0.9, 1.6, 1.2, 1.8	None
Glutathione S-T A	1K3Y†, 1K3L†	1.3, 1.5	SHG
Carbonic anhydrase II	3KS3, 3MWO, 2ILI	0.9, 1.4, 1.1	None

TABLE 3.2: The proteins and set of structures used to establish the docking and clustering procedures for the water placement method. *Structures that have been determined by neutron diffraction. †Structures where multiple chains from the same crystal structure have been used to validate ordered water molecules.

Each of the proteins in Table 3.2 were structurally aligned and consensus water sites were identified using the statistical programming language R [179]. Using a 15Å×15Å×15Å cube to define each binding site, 185 distinct water molecules were identified. Of these

water molecules, only 92 had been identified at least twice by experiment. Observing less than half of experimentally determined water molecules in at least two structures highlights the uncertainty regarding crystallographic water positions and underlies the need for caution when validating a water prediction method.

To test WaterDock on an independent data set, we chose 14 structures of OppA bound to different KXK tri-peptides (see Appendix, Table A.1). The data set was primarily chosen because the same test set was used for a recent water prediction method called AcquaAlta[169]. Doing so allows a direct comparison of the two methods. In addition, the structures have been determined to a high resolution and the ligands have varied water distributions around the side chain of the central amino acid[54].

3.3 Results

The minimum distance of each docked water molecule from a crystallographic or molecular dynamics (MD) water molecule was computed in order to assess how placement prediction error depended on the water position's Vina score. In particular, we sought to find a score cutoff that identified well-determined sites by comparing the predictions to a random placement of water molecules. Figure 3.1 shows how each Vina score has an error distribution associated with it and how the median and the range of the error distributions decreases for more negative (favourable) scores. In particular, as the scores increase (become less favourable), the distributions tend to the error distribution from the random placement model. It is apparent that the lower the Vina score, the closer the agreement with crystallographic water locations.

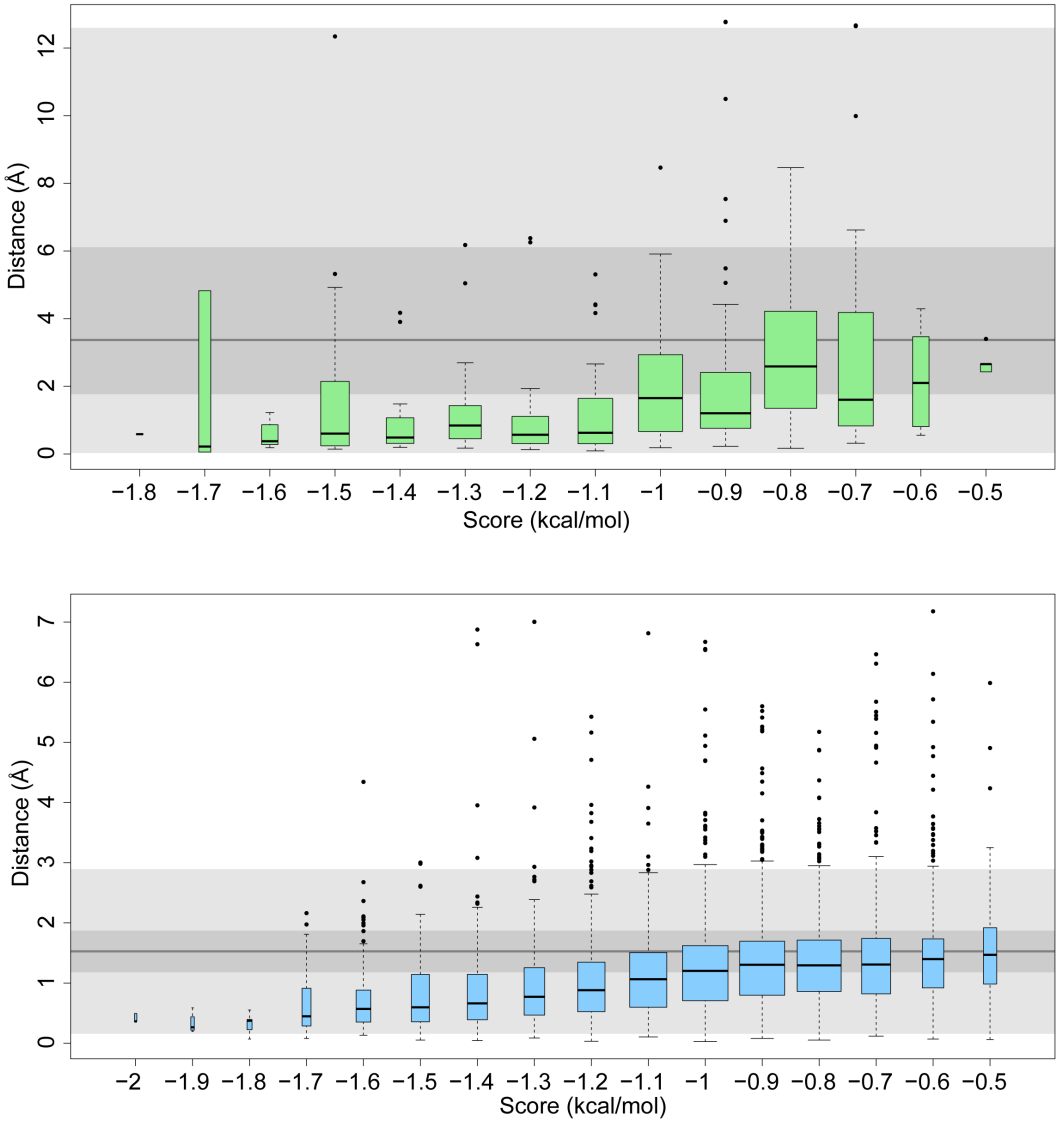


FIGURE 3.1: Box-plot summarizing the Vina score (in kcal/mol) versus the minimum distance (in Å) between the prediction and a crystallographic water (above) and MD water (below) from the data set in Table 3.1. Each box's lower and upper limits are at the 25th and 75th percentiles. The solid black line within each box indicates the median and the width of each box is proportional to the square root of the number of data points. Outliers are shown as black dots and are defined by points outside 1.5 times the interquartile range. For comparison, the results from a random placement of water molecules are shown by the grey background box (light grey represents the whiskers, darker grey represents the 25th and 75th percentiles and the darkest grey line represents the median). The accuracy of the placement increases with a more negative score and the median distance of the predicted sites is consistently lower than the median of the random placement method for scores less than -0.5 kcal/mol.

When predicting water locations in the X-ray crystal structures of Table 3.1, the error distributions were always better than the error distribution from the random model.

During the MD simulations, large numbers of water molecules filled the cavities. This meant that placing a water molecule at random within the cavity has a much greater chance of being near a simulated water molecule. While prediction error was also reduced, outperforming the random model provided a more stringent test. As a result, a cutoff of -0.6 kcal/mol was chosen by inspection as the minimum acceptable score of a predicted water molecule.

Using 7 crystal structures that had been resolved multiple times (see Table 3.2), different docking and clustering protocols were experimented with in order to find the method that predicted the largest number of consensus water molecules for the fewest number of false positives. Below, we summarize the most accurate protocol while the results of additional tested docking and clustering regimes are shown in Table A.2.

We found that independently docking a water molecule 3 times into the binding site appeared sufficient to sample the configuration space of the water molecule, while docking only once did not. The exhaustiveness parameter in Vina determines how rigorous the docking search is and is roughly proportional to elapsed docking time. We found that setting this parameter to 20 significantly improved the accuracy of the subsequent clustering methods when compared to an exhaustiveness value of 10. Three independent docking runs with an exhaustiveness value of 20 was also very fast and took no more than 15 seconds to complete on a 2.33 GHz Intel Xeon quad-core processor.

Independently docking a water molecule 3 times with Vina generates a maximum of 60 binding modes. Many of the positions overlapped or were in close proximity to one another. Clustering the water positions is a time efficient way of producing a solvation map of the binding site from an ensemble of water positions. A number of different hierarchical clustering methods were experimented with, including complete linkage, single linkage and Ward's minimum variance method. Distance cutoffs of each clustering

method were varied to find the one that gave the best accuracy. The average position of a docked water molecule cluster was used as the predicted water molecule location.

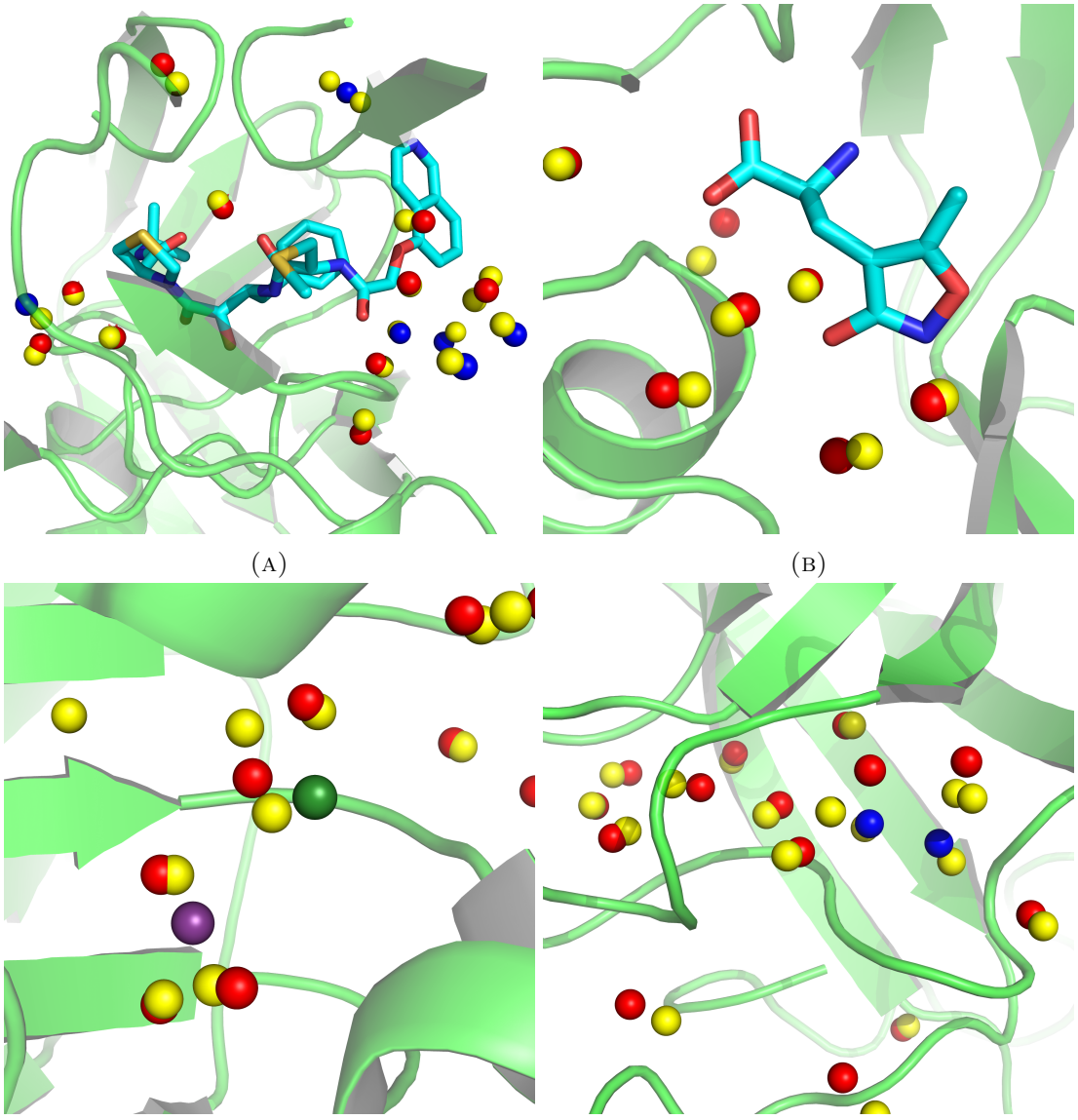


FIGURE 3.2: Examples from the data set in Table 3.2 that were used to develop the WaterDock method. Yellow spheres: predicted water sites, red spheres: water molecules observed in at least two experimental structures, blue spheres: water molecules observed in only one experimental structure. (A) HIV-1 protease bound to the inhibitor KNI-272. All 9 consensus water molecules and all 6 non-consensus water molecules are correctly identified. One non-consensus water molecule is in between two predictions, resulting in a false positive. This water molecule was resolved only in 3FX5 with a temperature factor of 42 $Å^2$, so the over-prediction may be due to the uncertainty in the water molecule's position. (B) GluR2 ligand binding core bound to AMPA. All water molecules within the binding site are correctly predicted. (C) Water around apo concanavalin A catalytic manganese and calcium ions (purple and dark green spheres respectively). Two false positives can be seen, with one over-coordinating the calcium ion. (D) Apo trypsin, showing good agreement between predicted and experimental water molecules.

Max. error = 1.5 Å

Max. distance of experimental water from protein	Consensus waters predicted (%)	False positives (%)	Mean error (Å)
3.0	88	24	0.69
3.3	81	24	0.69

Max. error = 2.0 Å

Max. distance of experimental water from protein	Consensus waters predicted (%)	False positives (%)	Mean error (Å)
3.0	94	16	0.77
3.3	88	16	0.78

TABLE 3.3: The performance of the final WaterDock method on the second validation set.

The most accurate clustering method was found to be with 2 rounds of single linkage clustering with different distance cutoffs. The results are summarized in Tables 3.3 and 3.4. The first clustering round used a distance cutoff of 0.5 Å and was designed to remove the most overlapping sites and to reduce the chaining of clusters in the second docking round. The output was clustered again with a distance cutoff of 1.6 Å. While these distance cutoffs were established empirically so as to maximize accuracy, the second clustering cutoff coincides with the effective van der Waals radius of a water molecule [180].

Using a maximum placement error of 2 Å the final WaterDock method identified 87% of consensus water molecules within 3.3 Å of the protein. The distance of 3.3 Å was chosen from the water-water radial distribution function so as to define the first hydration shell [181]. While used previously to validate water placement error [174], 2 Å is a lenient error cutoff. Yet, out of the 80 consensus water molecules correctly identified, only 8 were over 1.5 Å away from the experimental position and 54 were within 1 Å of a consensus water molecule. When only tightly bound water molecules (within 3 Å of the protein) were considered, WaterDock predicted 94% of these consensus water molecules.

Given that only protein-water interactions and not water-water interactions were used to generate the initial ensemble of positions, it is encouraging that WaterDock was able to predict the vast majority of consensus water sites. Even in examples that contain a

complex network of water molecules, such as ribonuclease A, and carbonic anhydrase II, WaterDock was still able to predict 80% of the consensus sites (see Table 3.4). It is clear therefore, that the protein is the most important factor in determining a water molecule's position. However, the omission of water-water interactions was likely to be responsible for some of the errors. In a few cases, an experimental water site was found to lie between two predicted locations (see Figure 3.2), resulting in a false positive. In examples such as ribonuclease A, concanavalin A and carbonic anhydrase II, it was found that water-water interactions were very subtle and consensus sites were observed to be slightly displaced with respect to the WaterDock predictions, possibly to accommodate and interact with another water molecule.

Water-water interactions could be included in the WaterDock method if a second sampling procedure, akin to the JAWS method[156] could switch the predicted sites on and off. We also considered sequentially docking a water molecule into a cavity to account for water-water interactions. However we found that the point at which to stop docking was ambiguous and that subsequent predictions were biased to regions near previous predictions. Importantly, neither of these methods adapt the positions of water molecules to optimize both the protein-water and the water-water interactions. A second energy minimization step would be required to achieve this. Given the high accuracy and speed of the current method, we felt these improvements were unnecessary.

Protein	Consensus waters	Predicted consensus waters	False positives	Water molecules predicted
HIV-1 protease	9	9	2	18
Ribonuclease A	10	8	3	20
GluA2	15	15	3	20
Trypsin	14	12	2	17
Concanavalin A	17	13	4	21
Glutathione S-transferase A	13	12	3	19
Carbonic anhydrase II	15	12	4	18
Total	93	82	21	133

TABLE 3.4: The individual protein results using the final WaterDock method for a maximum error of 2Å. The number of correctly predicted non-consensus water sites can be calculated by finding the difference between the number of water molecules predicted and the sum of the predicted consensus waters and false positives.

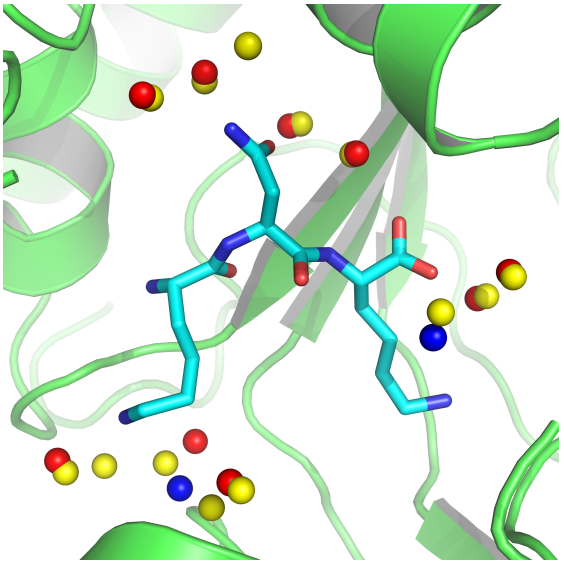


FIGURE 3.3: An example from the test set used to validate WaterDock. The tripeptide KNK is shown bound to OppA, PDB code 1B5I. Red spheres: crystallographic water molecules; blue spheres: water molecules seen in other related structures; yellow spheres: WaterDock predictions. All water molecules are correctly predicted with two false positives.

Using the OppA test set shown in Appendix Table A.1, AcquaAlta reported predicting 66% of the water molecules that bridged the interaction between the ligand and the protein to a maximum error of 1.4 Å. Using the same maximum error, WaterDock predicted 87% of the crystallographic water molecules. When the results were visually inspected, eleven additional predictions were found to be within 2.0 Å of crystallographic water molecules that made the same interactions with the ligand and protein. When these water molecules were included in the analysis, WaterDock identified 97% of the crystallographic water sites with a mean error of 0.68 Å. WaterDock predicted on average 1 water molecule per structure that was not seen experimentally. The false positive rate was not reported for AcquaAlta.

3.4 Discussion

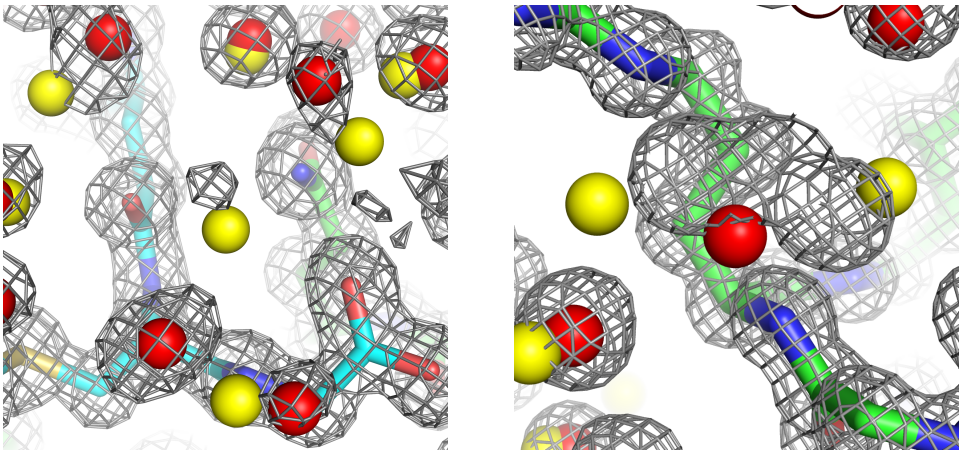
In the previous section, fixed distance cutoffs were used to classify WaterDock predictions as either "correct" or "incorrect". It is important to acknowledge that this approach has been used to simplify the error analysis, and in reality, the uncertainty regarding the true location of water molecules is more complex.

In crystallography, the degree of spread about the peaks in electron density represents the uncertainty in an atoms location. Structural heterogeneities between the molecules in the crystal, dynamic flexibility and thermal motion all contribute to this spread. Temperature factors, also known as B-factors, for each atom are proportional to isotropic variance of the electron density, which are readily interpreted as the mean square displacement of an atom from its average location. Therefore, the B-factors of water molecules can, in principal, be used to assess the accuracy of predicted water locations in a manner that reflects the inherent error in the crystallographic position. Anisotropic displacement parameters, which are provided for structures resolved to a high resolution

could also be used. However, such an error analysis may suffer from inconsistencies in the assignment of B-factor, which depend on the resolution of the structure and refinement method used by the crystallographer[182]. Hence, different standardisation strategies are commonly utilised when comparing B-factors from multiple crystal structures[183,184]. As standardised B-factors may be negative[53], they can no longer represent as mean square displacements.

Water molecules are often added to a protein-crystal structures in the latter stages of the refinement procedure[63]. As a result, water molecules may be erroneously included to reduce the average residual between the calculated structure and electron density, or simply may be missing. The significant discrepancy between water sites in crystallography and NMR[152] prompted us to validate WaterDock on waters observed in multiple structures. The false positive rate was estimated using all distinct water sites, whether consensus or not. Evaluating predictions against the crystallographic electron density is a more detailed analysis that, while free from refinement idyosyncrasies, is complicated by the fact that the concentration of other solutes in the crystal may be unknown and that electron densities are not reported for all structures in the PDB. Preliminary analysis suggests that our reported false positive rate may be an over estimate in some of the proteins (see Figure 3.4).

At thermal equilibrium, our ability to ascertain microscopic detail – such as collection of atomic coordinates – of a system is characterised by the Gibbs entropy. The location of water molecules in proteins is no exception, so we should expect the position entropy of a single hydration site, denoted $S_1^{\mathbf{r}}$, to be the ultimate limit of accuracy of any water placement method. The relationship between the mean squared error (MSE) of an estimator and Shannon entropy in Equation 2.22 in Chapter 2, applies equally for water placement error and $S_1^{\mathbf{r}}$, owing to the proportionality between Shannon and Gibbs entropy. Thus, when predicting the location of water molecules in protein structures



(A) (B)

FIGURE 3.4: Comparing crystallographic electron density maps against WaterDock predictions for (a) glutathione s-transferase and (b) concanavalin A. Yellow spheres: WaterDock predictions; red spheres: consensus water sites; green and blue sticks: protein and ligand respectively; grey mesh: electron density at one standard deviation above the background density. In (a) and (b), regions of electron density that may correspond to hydration sites are shown. In (a), the structure (PDB code 1K3Y) was resolved to 1.3 Å. Due to the significant overlap with consensus water sites, crystallographic waters are not shown. Although the structure in (b) (PDB code 1NLS) was resolved a resolution of 0.94 Å, there are large regions of electron density that are unaccounted for. The central consensus water at the centre of (b) was not present in the structure, and was inferred from the other apo concanavalin A structures from Table 3.2.

sampled from the Boltzmann distribution, assumed here to be locally Gaussian about the water molecule's mean position, one has

$$\mathrm{MSE}_{water} \propto \frac{1}{2\pi e} \exp\left(\frac{2S_1^{\mathbf{r}}}{k_B}\right). \tag{3.1}$$

The integral theory 3D-RISM can be used to predict not only the location of water molecules in protein-ligand complexes, but also the individual entropies of the water molecules. In agreement with Equation 3.1, a strong correlation was recently reported between placement error and entropy of water sites as predicted by 3D-RISM [161]. Clearly, error is an inescapable part of any method that predicts the discrete locations of water molecules; analyses based on distance cutoffs, B-factors or electron densities can

assess this error to varying degrees of validity, although each one is subject to its own limitations.

3.5 Summary

In this Chapter we have shown that freely available ligand docking tools can be appropriated to predict the location of hydration sites in protein structures. The simplicity of our procedure, which we call WaterDock, is the key to its speed; water molecules are docked into a structure using AutoDock Vina, low scoring sites are removed and the rest are clustered. The centroids of the clusters are the predicted water sites. Our method was validated against high-resolution crystal structures, neutron diffraction data and molecular dynamics simulations. In particular, we assembled a data set of high resolution protein and protein-ligand structures that had been determined at least twice. We found that WaterDock was able to predict 87% of consensus water sites with a mean error of 0.78 Å. Using 14 structures of OppA bound to lysine-X-lysine tripeptides, WaterDock predicted 97% of the ordered water molecules, with on average 1 false positive per structure. The key advantages of the approach we present here is that it takes only a few seconds to apply yet is able to maintain a high degree of accuracy.

Chapter 4

Predicting the role of water in protein-ligand binding

4.1 Introduction

When a ligand binds to a protein, water molecules vacate the binding site to accommodate the molecule, while others remain which bridge the interaction between the ligand and the protein. Water-mediated binding is so common that a study of 392 protein-ligand complexes found that 85% had at least one water molecule which bridged the interaction between the ligand and the protein[53]. The number and distribution of bridging water molecules can change for each new ligand, presenting a challenge for fast ligand binding mode prediction. As discussed in Chapter 1, Section 1.4.2, a popular strategy in rational drug design is to modify a ligand so that it displaces an ordered water molecule into the bulk solvent. But as targeted water displacement can fail, predicting which water molecules can be displaced by the ligand, and which remain to bridge the interaction is not only of concern to protein-ligand docking, but also to structure-based drug design.

Traditionally, ordered water molecules were ignored in ligand docking studies and ligands were docked into desolvated binding sites[185]. Presently, there are a number of docking protocols that include explicit water molecules and claim to improve accuracy in many cases[186–190], typically involving multiple docking runs with various different water positions. Water positions are typically selected from so called, "conserved" sites, which are hydrated in multiple structures of the protein bound to different ligands. It has also been reported that including water molecules may hamper efforts to predict a ligand's correct binding mode[191]. In addition, the chemical diversity of the hits from a virtual screen can either be reduced or increased, depending on which water sites have been retained[148,192].

Rather than preselecting which water molecules to keep in the binding site, some docking procedures – although different in implementation – involve switching water molecules "on" and "off" during docking[193–195] at additional computational expense. Some methods can simultaneously dock ligands and water molecules[175,196], where the method by Forli[175] docks ligands with water molecules attached to their hydrogen bonding groups.

A number of approaches have used the structural features of a water molecule's environment to predict whether it will be displaced or not without any prior knowledge of the ligand. Using a K-nearest neighbours genetic algorithm, Consolv reported 75% accuracy when predicting whether binding site water molecules would be displaced or not[197]. However, as Consolv used crystallographic temperature factors as structural descriptors, it cannot be applied to predicted water sites, such as with WaterDock as developed in Chapter 3. Amadasi and co-workers have combined the HINT forcefield[198] with the Rank score[199] to classify water molecules into two broad categories; conserved/-functionally displaced and sterically displaced/missing[200,201]. Their first study correctly classified 76% of the water molecules tested while their second study reported a classification accuracy of 87%. Their analysis included weakly bound water molecules, which

were a maximum of 4 Å away from the protein. Another method called WaterScore used water molecules within 7 Å of the bound ligand in protein-ligand binding sites[202]. Using multivariate logistic statistical regression, WaterScore reported 67% accuracy in classifying displaced and conserved waters, although water molecules that were displaced because of steric clashes with the ligand were not included in their analysis. Barillari et al. used the computationally expensive double-decoupling method to calculate the binding energies of 54 water molecules in protein-ligand complexes[116]. They found that water molecules that could be displaced by a ligand were on average less strongly bound than conserved water molecules by 2.5 kcal/mol.

Despite the positive strides that have been made in understanding the role of ordered waters, no single method is able to answer how displaceable a water molecule is, and what is it likely to be displaced by. Addressing these questions becomes even less clear when there is limited experimental knowledge of a binding site's solvation structure and one has to rely on theoretically predicted water positions. Here, we develop a conceptually straightforward method that can be used in conjunction with our WaterDock procedure to predict whether molecules are likely to be conserved or displaced after ligand binding. We also develop a model to predict the probability that predicted water molecules will be displaced by polar or non-polar groups.

4.2 Methods

As a rapid placement tool, the WaterDock method facilitates the development of empirical classification models as well as a large scale study into the statistics of water displacement. As described in more detail below, water molecules were predicted in the binding sites of the Astex Diverse Set[203] of protein-ligand complexes after the ligands had been removed from the structures. By overlaying the ligands back into the hydrated cavities,

we studied the statistics of hypothetically "displaced" water molecules. By developing novel water scores using data mining and heuristics, and a machine learning method, we developed probabilistic models to predict the role of water in protein-ligand binding. We developed probabilistic models rather than discrete classifiers because whether a water molecule is displaced or not depends on the size, type and scaffold of a ligand. Classifying a water molecule as either always displaceable or only conserved was deemed an oversimplification.

4.2.1 Establishing a water binding energy score

Using the double decoupling method, Barillari et al. calculated the absolute binding free energies of 54 water molecules from 35 ligand-protein complexes[116]. The data set was made up of 6 proteins and 11 conserved hydration sites. They found that conserved water molecules had statistically significant lower binding energies than displaceable water molecules. Therefore, we trained a water free energy estimator, with the aim to similarly distinguish conserved and displaceable water molecules.

We considered Barillari et al.'s data ideal to develop a water scoring function because of the size of the set, the diverse range of proteins and the consistent manner in which the binding energies were calculated. Rather than create a new scoring function from scratch, we opted to apply the Vina and AutoDock 4 scoring functions because of their free availability. Each of the 54 water molecules were initially scored using the scoring functions from Vina and AutoDock 4 and correlations with R^2 values of 0.01 and 0.31 respectively were found. It is interesting to note that while Vina is able to accurately dock water molecules, it has no predictive power on the binding free energies. If rigorous free energy calculations were used to dock and score ligands, accuracy in one would directly correspond to accuracy in the other. However, as both fast docking procedures

and scoring functions do not use rigorously calculated free energies, this correspondence cannot be guaranteed.

Due to the poor correlation in the Vina and AutoDock 4 scoring functions, we sought to find which combination of components in the scoring functions had the best predictive ability. We wished to do so in a purely data driven manner, without imposing any bias as to which sort of interactions we thought would be important for water, to compare and contrast to our heuristic approach to a water scoring term as described in Section 4.2.2.

AutoDock 4 has 5 terms in its scoring function, while Vina has 6. Excluding Autodock 4's and Vina's rotatable bond terms and Vina's hydrophobic term, both scoring functions have 4 terms each that are applicable to water. There are 255 unique combinations of these terms, with the largest having 8 terms: 4 from Vina and 4 from AutoDock 4. Linear regression was used to fit each of the 255 combinations of terms results in 255 candidate water scoring functions. We note that because each model has fewer explanatory variables than the number of response variables, the solution to each linear model is unique. As a vast majority of these models were overfit, we wanted to extract from this pool of water models the one with the best balance between accuracy and simplicity. Thus, each of the models were ranked by their Akaike information criterion (AIC)[204]. As discussed in Chapter 2, Section 2.4, the AIC is a information theoretic measure for the goodness of fit of a model, which penalises models for the number of parameters they contain, the preferred model being the one that minimises the AIC. The top 30 models with the lowest AICs were then selected for an extensive cross validation study to further assess the trade-off in model complexity and accuracy.

4.2.2 Development of hydrophilic and lipophilic scores

By analyzing 10,837 surface bound water molecules in 56 high resolution crystal structures, Kuhn et al. established the individual hydration propensities for each amino acid atom type[205]. They determined the propensities by dividing the total number of water molecules that hydrated an atom by the number of surface exposed occurrences. Building on their work, we created a hydrophilicity model and a lipophilicity model intended to encapsulate the local chemical environment of a water molecule. These models were intended to be distinct from the water energy model previosly described, as they are heuristic knowledge-based potentials as opposed to a fitted forcefield model. The hydrophilicity model is a distance weighted sum of the propensities from all the atoms within 4 Å of a water molecule and is given by:

$$U_H = \sum_{i}^{N} p_i \exp\left(-\frac{r_i}{d_0}\right), \tag{4.1}$$

where N is the number of protein atoms within 4 Å of the atomic position, r_i is the distance (in Å) of protein atom i to a water molecule, p_i is the hydration propensity of atom i and d_0 is the distance scale of the interaction, set at 1 Å. We chose the weighting function because previous works have suggested that hydrophobicity decays exponentially with distance[206], albeit at longer length scales, and is used by the HINT forcefield[198]. The hydration propensities of cofactor atoms were assigned the same value as the most similar protein atom. Because of the high magnitude of ion hydration free energies, ion hydration propensities were assigned the same as the highest value in the Kuhn data set. For the lipophilic score, we chose the same form as Equation 4.1 and it is given by

$$U_L = \sum_{i}^{N} l_i \exp\left(-\frac{r_i}{d_0}\right), \tag{4.2}$$

where the terms are as before except l_i which is the carbon propensity of protein atom i. As atomic carbon propensities have not been established as they have been for hydrophilicity, for simplicity, we set all carbon atoms a propensity score of 1 and all other atom types a score of 0.

4.2.3 Finding displaced water molecules retrospectively with WaterDock

The Astex Diverse Set contains 85 high-resolution crystal structures of pharmacologically relevant ligand-protein complexes[203]. The ligands are drug-like and have a diverse range of scaffolds. Importantly, the electron density of the ligands in the crystal structures accounts for all parts of the ligand, leaving little ambiguity over the binding mode and the locations of the ligand atoms. This makes the Astex Diverse Set appropriate data to investigate what types of ligand atoms displace WaterDock predictions of apo water molecules.

Ligands and water molecules were removed from the binding sites and cofactors were retained. Water sites were predicted in the binding site using the WaterDock method as described in Chapter 3. A predicted water molecule was classified as conserved if it was seen within 1.5 Å of a water molecule seen in the crystal structure of the protein-ligand complex. Predicted water molecules that were not within 1.5 Å of a crystallographic water molecule but within 1.5 Å of a ligand atom were classified as displaced. The distance cutoff was chosen as this represents an acceptable water prediction error, being within the van der Waals radius of a water molecule[180].

4.2.4 Development of a probabilistic water classifier

We expect that the displacement probability of a water molecule depended on a – potentially – non linear combination of the 3 structural descriptors (binding energy, hydrophilicity and lipophilicity) and that certain regions of parameter space would generally correspond to different classes of water molecule. Classification trees meet these requirements by recursively partitioning the parameter space such that each region defines a class. Classification trees are particularly well suited to our problem because the proportion of a class in a partitioned region can be readily interpreted as a conditional probability. However, because of a tree's hierarchical nature, small changes in the data can result in a different series of splits, making single classification trees unstable. The method of bootstrap aggregation – known as bagging and discussed in Chapter 1, Section 2.3 alleviates this issue by fitting many trees to bootstrapped samples (sampling with replacement) of the data. The probability of a class is found by averaging the class proportions from each classification tree.

Using the free statistical language R with the package "rpart" [179], a bagged classification tree algorithm was written and trained on the predicted water positions in Astex Diverse Set to classify them as conserved or displaced. In addition, a second model was trained to classify displaced WaterDock predictions as displaced by hydrogen-bonding groups or by non-polar groups. To assess the accuracy of the models, we used leave-protein-out cross validation. This involved partitioning the Astex Diverse Set into a training set and a test set, where the test set comprised of all the water molecules from a single protein. The water molecules in the test set was classified by both models and the fraction of correct predictions were recorded. This process was repeated until all 85 proteins had been used as the test set. The accuracies quoted in the results are the mean accuracies from all partitions. This validation procedure was chosen so that the models were tested on structures that were distinct to the structures in the training set.

Term	Coefficient (kcal/mol)
Intercept	1.77
H-bond	-2.58

TABLE 4.1: The gradient and intercept of Vina's hydrogen-bonding term after refitting to the calculated binding energy of water according to Barillari et al[116].

4.3 Results

4.3.1 Water energy model from a data mining procedure

On Barillari's et al.'s data set of calculated water binding energies, it was found that a single term, the hydrogen bonding term from Vina's scoring function, had the lowest mean error in the cross-validation study, with an error of 1.7 kcal/mol. The standard error of the fit was 1.6 kcal/mol and had an $R^2 = 0.50$. For comparison, if the average calculated energy of the Barillari data set is used to predict each water molecule's energy, the mean error would be 2.5 kcal/mol. The coefficient and intercept of the re-weighted Vina hydrogen bonding term is shown in Table 4.1.

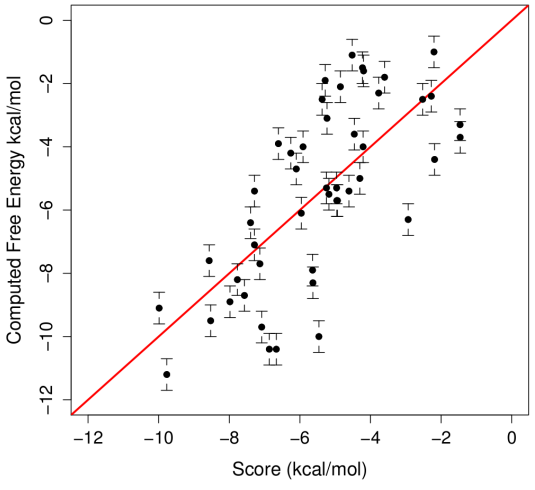


FIGURE 4.1: The re-weighted AutoDock Vina score applied to training set set of calculated protein-water binding free energies. The $y = x$ line is shown in red.

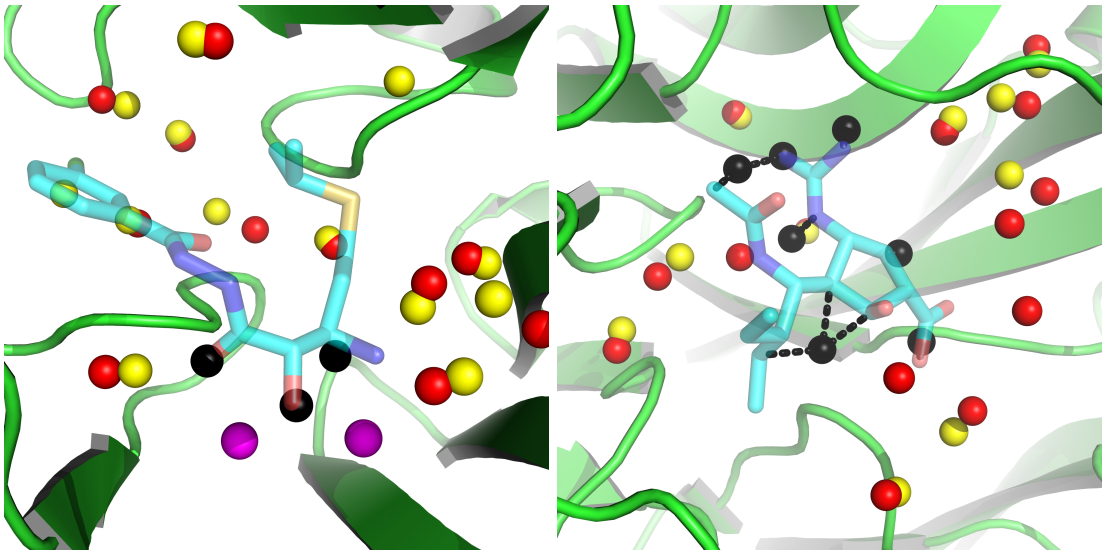
Vina's hydrogen bonding term is the sum over hydrogen bonding pairs[176]. For each pair, the value ranges from 1 to 0 and varies linearly with distance. The significant correlation

despite the simplicity of the model is likely to be due to a strong enthalpy-entropy compensation effect, where the number and strength of hydrogen bonds correlates with the lack of translational and orientational freedom of the water molecule.

4.3.2 Classifying the role of water

To develop a water classifier that was consistent with our water placement method, water sites were predicted in the Astex diverse data set after the ligands had been removed from the structures. By overlaying the ligands back onto the hypothetical apo solvation structure, we investigated the displacement statistics of our water predictions (See Figure 4.2). In total, 545 predicted apo water molecules were within 1.5 Å of water molecules seen in the crystal structure of the protein-ligand complex and so were classified as conserved. Also, 459 predicted water molecules were classified as displaced as they were within 1.5 Å from a ligand. Of these displaced water molecules, 216 were displaced by polar groups and 243 were displaced by non-polar groups.

As described, the water energy model was trained on the 54 water molecules in Barillari et al.'s study, which had binding energies calculated using thermodynamic integration[116]. In that study, they found that displaced water molecules were less strongly bound than conserved water molecules, with an average difference of 2.5 kcal/mol. Likewise, the scores of the hypothetically displaced water molecules in our displacement study were also predicted to be less strongly bound, with a difference of 2.0 kcal/mol compared to the hypothetical conserved waters. It is encouraging that the simple model captured this relevant feature for water displacement. As our study classed water molecules as "conserved" if they were not displaced in a single complex, while Barillari et al. as water molecules for which there is no evidence of displacement, one could expect our score to predict a wider difference if the latter criterion is applied.

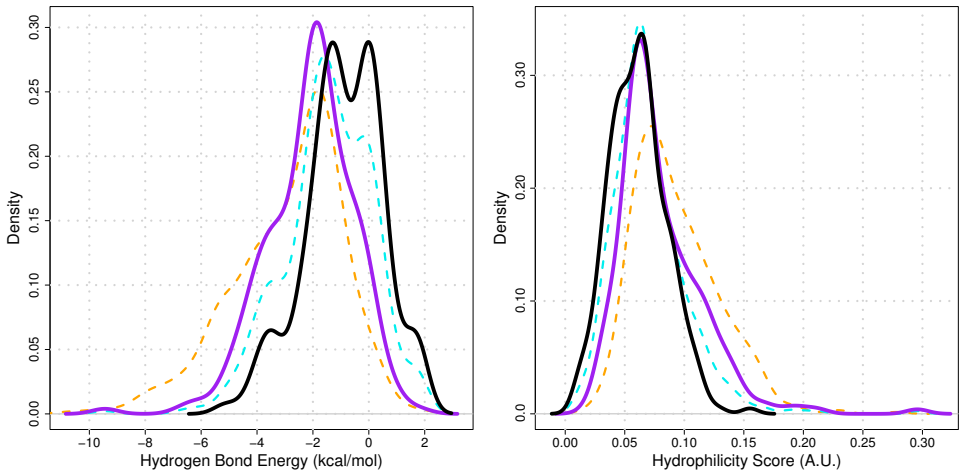


(A) (B)

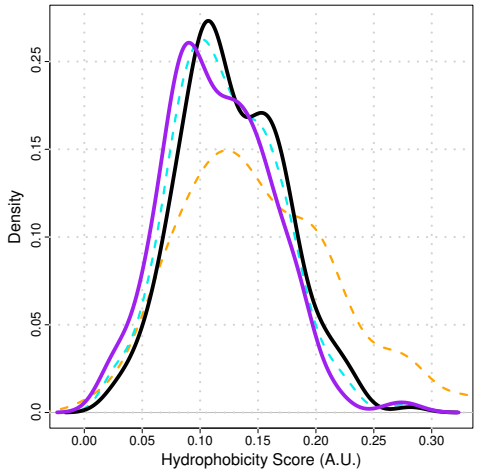
FIGURE 4.2: Two examples the retrospective displacement study. (A) Human methionine aminopeptidase-2 and (B) neuraminidase bound to inhibitors (blue transparent sticks), PDB codes 1R58 and 1L7F respectively. Yellow spheres: water sites predicted in the absence of the ligand; black spheres: predicted water sites that overlap with the ligand; red spheres: crystallographic water molecules seen in the protein-ligand complex, purple spheres: manganese ions. Predictions that correspond to water molecules seen in the crystal structure are considered to be conserved and water molecules that overlap with the ligand are considered to be displaced. Multiple ligand atoms are considered to displace water molecules, and in (B), ligand contacts within 1.5 Å of predicted waters are highlighted with black dotted lines

Using the re-weighted Vina hydrogen bond term, the hydrophilicity model and the lipophilicity model as descriptors in a probabilistic machine learning classifier, water molecules were predicted to be either being displaced or conserved. Using leave-protein-out cross validation (as described in Methods, Section 4.2.4), 75% of the WaterDock predictions are correctly classified as either conserved or displaced when the class with the highest probability was used for the prediction. Similarly, when waters predicted to be displaced by WaterDock are classified as being displaced by a polar group or by a non-polar group, 80% of the WaterDock predictions are correctly classified in cross validation. Table 4.2 shows that there appears to be a slight bias for predicting water molecules to be displaced rather than conserved.

One benefit of using a probabilistic classifier is that the certainty of a prediction is



(A) (B)



(C)

FIGURE 4.3: The distribution of energies of water molecules displaced by polar groups (purple) and by non-polar atoms (black) for the hydrogen bond score (A), hydrophilicity score (B) and the lipophilicity score (C). Overlaid are the distribution of energies of conserved water molecules (dashed orange) and displaced water molecules (dashed cyan). Water molecules that are conserved and functionally displaced tend to be in more hydrophilic regions with stronger binding energies than water molecules that are displaced by polar groups.

naturally quantified. One would therefore expect that the higher the classification probability is, the lower the chance of misclassification. For both of our models, we found that classification probabilities of 0.80 or above correctly classified the water in 94% and 95% of cases in both models after cross validation.

In Figure 4.3, it seems counterintuitive that conserved WaterDock predictions are more likely to have a higher lipophilic score than displaced water molecules. This is due to

Model 1 correctly classified (%)		
Total	Conserved Waters	Displaced Waters
75	70	81

Model 2 correctly classified (%)		
Total	Displayed by polar group displaced	Displaced by non-polar group
80	82	79

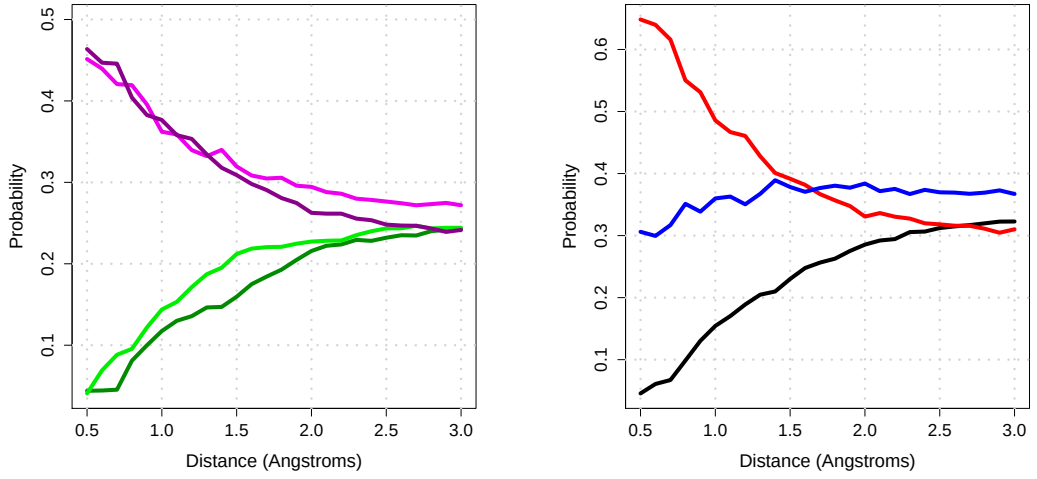
TABLE 4.2: The results of the models that classify water molecules as displaced or conserved and as displaced by a polar group and displaced by a non-polar group.

the fact that conserved water molecules tend to be more buried with higher numbers of protein contacts than displaced waters, which also explains the higher hydrophilicity scores and the stronger hydrogen bonds. The opposite is true when one compares WaterDock predictions that were displaced by polar groups to water predictions that were displaced by non-polar groups. Water molecules displaced by non-polar groups tend to reside in slightly more lipophilic and less hydrophilic environments and tend to make fewer and weaker hydrogen bonds.

It is interesting to note that even though Vina's hydrogen-bonding term was established using a data mining protocol and the hydrophilicity score was designed heuristically, both scores were strongly correlated with an R^2 of 0.72. These very different approaches have converged to describe a related property of water. Despite the high correlation, the combination of the two scores in the machine learning algorithm increased the classification accuracy by around 7% compared to when each term was fitted individually, suggesting that despite the high correlation, both terms contain sufficiently distinct information.

4.3.3 Ligand water displacement propensities

As well as predicting the role that WaterDock predictions play in ligand binding, we also investigated the propensities for ligand chemical groups to occupy the predicted water sites in the Astex Diverse Set. Given the good agreement between WaterDock's



(A) (B)

FIGURE 4.4: (A) The probability of finding a ligand atom, by group, at a given distance from a predicted water site. Light purple: hydrogen donors, dark purple: hydrogen acceptors, light green: aliphatic carbon atoms, dark green: aromatic carbon atoms. (B) The probability of finding a ligand atomic element at a given distance from a predicted water site. Red: oxygen, blue: nitrogen, black: carbon. Within 1.5 Å, ligand oxygen atoms are far more likely to occupy predicted water sites than any other element.

predictions and experimentally determined water sites, we expect these displacement statistics to be similar for water molecules seen in crystal structures.

Figure 4.4 shows the probability of finding ligand functional groups at various distances from hypothetically displaced water sites. For a given distance cutoff, each point can be considered as the propensity that a ligand atom will displace a water molecule. Hydrogen bond donors and acceptors were equally likely to displace predicted water molecules and were found to be around 9 times more likely to be within 0.5 Å of a water site than aromatic and aliphatic carbons. This indicates that it is important for water displacing ligand groups to replicate water's hydrogen bonding capacity. Interestingly, when the occupation probabilities were computed for ligand atoms, rather than atom functions, oxygen atoms were over twice as likely to be found within 0.5 Å of a displaced water site than nitrogen atoms. At 1.5 Å – the distance cutoff we previously used to define whether a water molecule was displaced or not – the displacement propensities of oxygen and nitrogen are roughly the same. In large part, this is due to the frequency and shape of

the chemical groups that contain oxygen atoms compared to those that contain nitrogen atoms. Elements which are only covalently bonded to one other non-hydrogen atom protrude more from a ligand than elements that are covalently bonded to at least two heavy atoms. These protruding elements can, due to steric reasons, more easily occupy partially buried water sites. For all of the 85 ligands in the Astex Diverse Set, the number of occurrences where the terminal heavy atom of a chemical group was a nitrogen atom (such as in primary amines or nitriles) were counted, as were the number of occurrences where oxygen was the terminal heavy atom (such as in alcohols or carbonyls). When adjusted for the total number of oxygen and nitrogen atoms, there were 4.6 times as many protruding oxygen atoms than there were protruding nitrogen atoms. Thus, based on ligand shape alone, one would expect oxygen to have a higher relative displacement propensity at shorter distances compared to nitrogen for this data set, as observed in Figure 4.4.

As the distance from a predicted water site increases further, the less one can consider a ligand atom to have displaced a water molecule. As a result, the propensities tend to the same value. Ligand atoms such as halogens, sulfur and phosphorous were not included in this study due to their small number in the data set.

From Figure 4.4, it is tempting to conclude that ligand modifications designed to displace a water molecule should always be made with an hydrogen-bonding group. However, in this study we have seen that many water molecules, depending on their local environment, are preferentially displaced by non-polar groups. Yet, since carbon is the most abundant ligand element in the Astex Diverse Set – and most organic compounds – the per atom displacement probability is significantly less for carbon than for polar atoms.

4.4 Summary

In this Chapter, we developed three water-specific scores that can accurately predict the probability of a water being displaced and conserved, as well being displaced by a polar group or non-polar group. One score in particular – selected to correlate with the binding free energies calculated using thermodynamic integration – replicated the differences between the binding energies of hypothetical conserved or displaced water molecules in the Astex Diverse Set of protein-ligand complexes.

Given their fast implementation and ability to capture useful information regarding the role of water in protein-ligand, these models, when used in conjunction with our placement method WaterDock, provide a starting point for the development of a novel water-based scoring function.

Chapter 5

Explicit water in empirical scoring functions

5.1 Introduction

Despite decades of research, predicting the affinity of a protein-ligand complex using a single snapshot of the bound state remains a notoriously difficult problem. Addressing this problem is of great concern to virtual screening, especially in the latter stages of a screening cascade, where protein-ligand docking and scoring functions are used to enrich a chemical library with active compounds and, more ambitiously, rank order series of ligands by affinity.

Over the years, progress in improving the accuracy of scoring functions for affinity prediction has been slow. For instance, in 2003, a survey of eleven scoring functions tested on one-hundred diverse protein-ligand complexes reported that the Pearson correlation coefficients of the predictions against experimentally measured affinities ranged from 0.42 to 0.70[207], while a similar survey conducted in 2009 with one-hundred and ninety-seven complexes and sixteen scoring functions reported correlation coefficients between 0.55

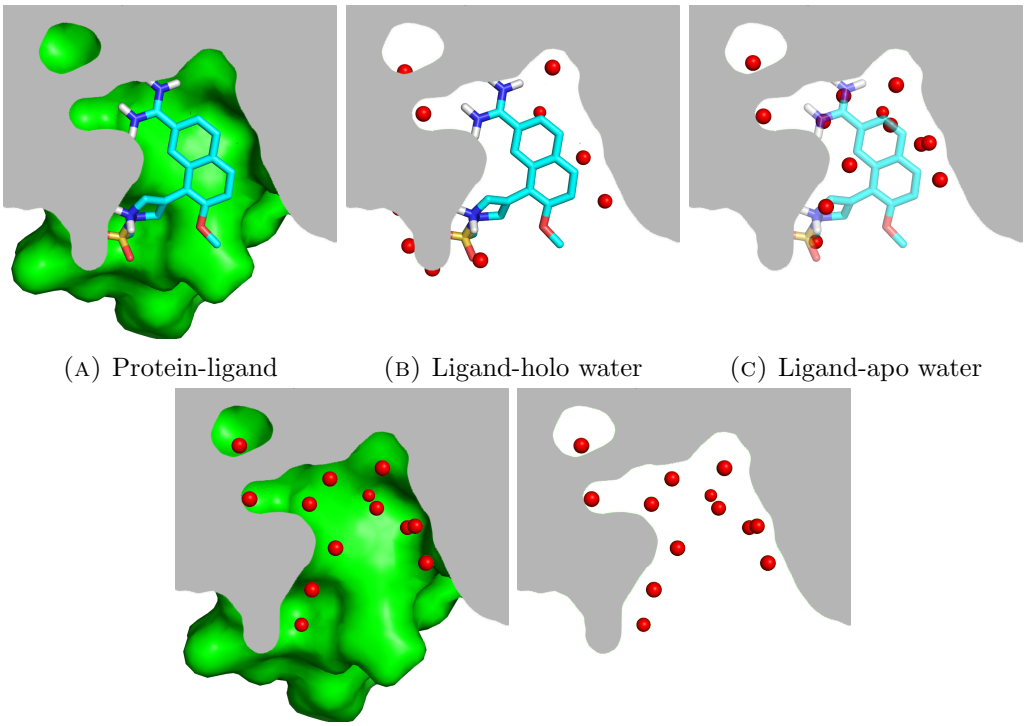
and 0.64[28]. It has also been found that accuracy is heavily protein-dependent; scoring functions with high enrichment rates on one protein may perform worse than random on another protein[25,26,31,33,194]. As most scoring functions either ignore water effects completely or use a type of implicit solvent model, it has often been suggested that a more rigorous treatment of water will increase their accuracy[56,208,209]. The rationale behind such arguments stems from the reported success from a number of affinity prediction methods that utilise explicit waters. The inclusion of bridging water molecules in protein-ligand binding interfaces has been reported to improve free energy perturbation calculations[62] and Poisson-Boltzmann and generalised Born surface area techniques[210,211], despite the fact that these latter methods have been designed to be used in the absence of explicit water molecules. The consideration of apo water molecules also forms the basis of protein-ligand affinity calculations using inhomogeneous solvation theory[119,120], which is exemplified by the WaterMap method[123,124] (see Chaper 1, Section 1.4.3).

To include structural water molecules in a scoring function, one must first know their location in the apo and holo three-dimensional protein structures. While methods that utilise molecular simulations represent the gold standard in accurate water placement predictions[153,155], they are too slow to be used for virtual screening, as the distribution of bridging holo waters would have to be recalculated for each new ligand. Our WaterDock method, discussed in Chapter 3, provides a possible solution: the method is fast enough to be applied to thousands of complexes in a matter of hours, and potentially accurate enough to be combined with a rapid affinity model. The water scores developed in Chapter 4 are adept at capturing salient features of waters in protein-ligand interactions, and can form the basis of a water-based scoring function. The aim of this chapter is to test whether a scoring function using apo and holo water locations prediction by WaterDock can improve rapid affinity estimates.

5.2 The model

What type of scoring function is best suited to incorporate water molecules? Scoring functions classically fall within three main categories: forcefield, knowledge-based and empirical[24,209]. Empirical scoring functions, (of which early pioneers include Krystek et. al.[212] and Böhm[213]) often feature as the most accurate models in independent tests[28,207,214]. They form the broadest category of scoring function as they use regression to relate a predetermined set of structural descriptors to experimentally determined binding affinities. The addition of water in an empirical scoring function thus requires the selection of water-based descriptors for both apo and holo waters. To test whether the addition of explicit water detail improves scoring accuracy, either such terms can be added to a current, well known scoring function, or a new scoring function can be created specifically for the analysis. Here, we have opted for an approach between the two, using descriptors that are based on previous models as well as our own. This allows for greater flexibility in the descriptors we choose and facilitates the consistent scoring of protein, ligand and water interactions. Scoring is performed using our own software written in Python and R[179].

In general, the binding free energy of a protein-ligand complex depends not only on the interactions between the protein and ligand, but also their interactions with the surrounding solvent. Water is not a passive medium in which binding takes place, but is an intrinsic part of the complex. In order to include explicit water effects into a tractable empirical model, some simplifying assumptions are necessary. First, we assume that the contribution holo waters make to the binding free energy depends only on their interaction with the ligand and are considered part of the protein. Second, we assume that the contribution apo waters make to the binding free energy depends on their interaction with the overlaid ligand *and* their environment prior to displacement, which includes both the protein and apo water-water hydrogen bonds. For example, we



(A) Protein-ligand (B) Ligand-holo water (C) Ligand-apo water

(D) Protein-apo water (E) Apo water-water

FIGURE 5.1: The interaction pairs that we consider in our empirical scoring function. The binding site cross-section is shown in grey. The explanatory variables for each pair are calculated using a subset of the interaction potentials outlined in Section 5.2.1. For each interaction pair, the sum over a particular interaction potential is used as an input into the model. For instance, the hydrogen bond potential (Equation 5.6) summed over all (A) protein-ligand donors and acceptor pairs forms one input, the hydrogen bond potential summed over all (B) ligand and holo water molecules forms another input, and so on. The interaction potentials used for pairs (A) to (E) are shown in Table 5.1.

expect that the affinity contribution an apo water bound in a polar environment makes is dependent on whether it is displaced by a polar ligand group or an apolar group, and that the contribution would be different if it was bound in an apolar environment. Third, apo water molecule locations are predicted using protein-ligand complexes after the ligands have been removed from the structure. This assumes that the binding site of the protein has not undergone a significant conformational change after ligand binding.

We approximate the binding free energy, denoted ΔG, of a protein-ligand complex as an arbitrary function of n interaction terms of the protein-ligand and water complex:

$$\Delta G = f(U_1, U_2, ..., U_n), \tag{5.1}$$

where U_i denotes a particular interaction term. The functional form of the model depends on the method of regression, which is discussed below in Section 5.2.2. To test whether the inclusion of explicit water detail improves the accuracy of scoring, we use three sets of interaction potentials: Set 1 uses only the descriptors for protein-ligand interactions, Set 2 the protein, ligand and holo waters, Set 3 the protein, ligand, holo and apo waters. For simplicity and consistency, we use only seven interaction types for the ligand, protein and holo and apo waters, which are described in the following section.

5.2.1 Interaction terms

Essential to the scoring of non-covalent interactions are terms that capture the Pauli repulsion of atoms that are close together and terms that capture the long range attraction due to electron dispersion. Together, they are commonly referred to as van der Waals (vdW) forces. Typical among forcefields for molecular simulations[215,216] and forcefield-based scoring functions[217,218] is the Lennard-Jones 12-6 potential, which captures both of these features. While asymptotically correct for attractive dispersion forces, its hard core potential is *ad hoc*, originally introduced for computational efficiency[219]. Interestingly, the Lennard Jones 9-6 and 8-4 potentials can also be found in popular scoring functions[207], suggesting there is a great deal of freedom in the functional form of the potential as long as the basic repulsion and attractive forces are present. Popular empirical models have also taken a 'shape-based' approach to vdW interactions[176,220–222], seeking to reward high surface area contact between the ligand and protein whilst minimising

volume overlap. This has been achieved using a sum of Gaussian distribution function based terms[176,220,221], or in the case of X-Score[222] with a Lennard Jones 8-4 potential with the same minimum energy-well depth for each atomic pair. We take a similar approach to these shape-based interactions, and use the following Morse potential

$$U_M = \sum_{i,j} \left[1 - \exp\left(-\alpha(r_{ij} - R_{ij})\right)\right]^2, \tag{5.2}$$

where r_{ij} is the distance between the atoms i and j, R_{ij} is the sum of their van der Waal radii and α is the parameter that sets the width of the potential. The above Morse potential contains fewer free parameters than a sum of Gaussian terms, and has a similar shape to the Lennard-Jones 6-12 potential. To reproduce the scale of the Lennard-Jones potential, we set $\alpha = 6/R_{ij}$[223]. However, unlike the Lennard-Jones 6-12 potential, U_M tends to zero at a faster rate for an increasing r, thereby rewarding larger atomic distance less, and has a softer hard-core potential. For an atomic pair, the above potential has a minimum of -1, and we truncated U_M at a value of 2 for small r, so that a 'bad' fit is penalised more than a 'good' fit is rewarded. The vdW radii of the atoms were taken from Bondi[224], which were derived from small molecule crystal structures. These radii are consistent with vdW radii derived from statistical atomic contact data from protein crystal structures[180,225] with the benefit of having radii for elements such as halogens, which are common features of drug molecules but not found in proteins.

As discussed at length in Chapter 1, hydrophobicity is often reported to be one the major driving factors for molecular association[67,78]. Indeed, the buried apolar surface area of a ligand when complexed with a protein has been found to significantly correlate with binding affinity[104,226]. In addition, Kellogg and co-workers have made hydrophobic "interactions" an integral part of their HINT (for hydropathic interactions) model[227], by

using computed atomic logPs to score interaction pairs. We have sought to replicate this information by employing the lipophilic and hydrophilic interaction terms we originally developed for our water classifiers. Although described in Chapter 4, they are described here again for completeness. The lipophilic interaction potential, denoted U_L, is simply a sum of carbon contacts weighted by their distance, so that

$$U_L = \sum_{i,j} l_i \exp\left(-\frac{r_{ij}}{d_0}\right), \tag{5.3}$$

where l is equal to one for carbon atoms and equal to zero for all other elements. Thus, this term is high when an apolar ligand atom is placed in an apolar protein environment. The d_0 is a constant set to equal 1 Å and is included set to scale the interaction and to ensure to non-dimensionality of the exponent. Similarly, the hydrophilic term, U_H, is given by

$$U_H = \sum_{i,j} p_i \exp\left(-\frac{r_{ij}}{d_0}\right), \tag{5.4}$$

where p is the hydration propensity of the atom. For protein atoms, these were taken from the study by Kuhn et al[205]. For ligand atoms, the average values of the protein hydration propensities for each element were used. For simplicity, halogens were given the same hydration propensity as the average carbon value; an attempt to capture halogen bonding would require an escalation of complexity that is unnecessary at this early stage. The hydrophilic interaction term is high whenever a polar ligand group or water is placed in a hydrophilic environment, irrespective of the type of interaction with the protein. In the case of protein-ligand interactions, it is hoped U_H acts like an

implicit apo water field, to be compared to explicit water contributions to affinity. To account for hydrogen bonds, we use the same function found in X-Score, AutoDock Vina and others[176,222,228]. For a particular pair of atoms, the hydrogen bond strength B_{ij} is given by

$$B_{ij} = \begin{cases} 1 & \text{if } (r_{ij} - R_{ij}^X) \leq -0.7 \\ -10(r_{ij} - R_{ij}^X)/7 & \text{if } -0.7 < (r_{ij} - R_{ij}^X) \leq 0 \\ 0 & \text{if } (r_{ij} - R_{ij}^X) > 0, \end{cases} \tag{5.5}$$

where $R_{i,j}^X$ is the sum of the X-Score van der Waals radii of the atoms[222]. The total hydrogen bond score is then given by

$$U_B = \sum_{i,j} B_{ij}. \tag{5.6}$$

This hydrogen bond score is chosen among others due to its success in predicting the calculated binding affinities of water molecules as found in Chapter 4, and the ease of its implementation. Despite the fact that the orientations of the hydrogen atoms are ignored, X-Score is often reported as one of the most accurate scoring functions[28,207]. The interaction U_B is appropriately short ranged, with a maximum of 1 when the X-Score vdW radii overlap, and is zero for donor-acceptor pair distances greater than 3.4 Å.

In the same manner as many classic empirical and semi-empirical scoring functions[213,217,220], we use a torsional score U_T, given by

$$U_T = N, \tag{5.7}$$

where N is the number of rotatable bonds in the ligand. The rotatable bonds are first identified using AutoDockTools (ADT)[177] and adapted using our own software written in Python, so that bonds between aromatic rings and carboxylic acids or imidamide groups (such as in benzamidine) are not counted as rotatable. Bonds between two aromatic rings are treated as rotatable, which is the default in ADT.

While the above terms will be applied to score protein, ligand and water interactions, another term is required to help distinguish whether an apo water overlaps with the binding mode of the ligand. In Chapter 4, we classified a water molecule as being "displaced" if it was within 1.5 Å of a ligand atom. Here, we opted for an approach which accounts for the vdW radii of the ligand and water atoms, and for any uncertainty as to whether it has been displaced or not by using the following logistic function as a water displacement/ligand proximity potential:

$$U_D = \sum_{i,j} \frac{1}{1 + \exp\left(\beta(r_{i,j} - R_{i,j})\right)}, \tag{5.8}$$

where the point of inflection is the distance at which the van der Waals radii overlap and β is a constant that determines the steepness of the rise. Unlike the numeric value of U_M calculated between the ligand and water, which can also be used determine overlapping vdW radii, the logistic function is monotonic and therefore more suitable to determine the likelihood of water displacement. We set $\beta = 10$, so that the potential was approximately zero for distances larger than ~ 4 Å and one for distances less than

$\sim$ 3 Å. For each water molecule, we truncate the score to a maximum of 1, to indicate complete displacement. For a set of apo water molecules, functional displacement by the ligand is indicated by a high U_D and a high U_B or U_H with the ligand. In the same manner, water molecules displaced by apolar groups will have a high U_D and high U_L with the ligand. Similarly, the interaction of apo water molecules with the protein is also considered (see Figure 5.1 and Table 5.1).

Like AutoDock Vina's scoring function, we have only used hydrogen atoms to classify polar atoms as either donors or acceptors in hydrogen bonds; hydrogen atoms themselves are ignored when scoring. This was to simplify the set up of complexes for scoring, as optimising the hydrogen bonding network, for instance, of the apo water molecules is a non-trivial task, and requires dependency on other software. Despite this simplification, Autodock Vina has been shown to compare favourably to AutoDock 4 – which scores hydrogen atoms explicitly – in virtual screening tests with HIV-1 protease[229]. In addition, the water models in Chapter 4 were sufficiently accurate despite ignoring the orientations of hydrogen atoms.

A notable omission from our descriptors is an electrostatic interaction term for each of the interaction pairs shown in Figure 5.1. A vital component in all forcefield-based scoring functions, we expect that the general features of this interaction are already captured in our hydrogen bonding and hydrophilic interaction terms. Crucially, in an initial analysis we found that AutoDock 4's screened Coulomb potential did not correlate at all ($R^2 = 0.004$) with the experimentally determined binding affinities of the protein-ligand complexes of the 2009 edition of the CSAR data set. In addition, a recent study showed that improving the partial charge model of AutoDock 4's scoring function does not significantly improve ranking and docking performance[230]. Results such as these have lend credence to the argument that while electrostatic interactions are critical

for specific interactions, Coulombic potential energy has little explanatory power over protein-ligand binding affinities in a diverse data set[231].

	Interaction Pairs				
Potential	Protein-ligand	Ligand-holo water	Ligand-apo water	Protein-apo water	Apo water-water
U_M	● ● ●	● ●	●	●	
U_L	● ● ●	● ●	●	●	
U_H	● ● ●	● ●	●	●	
U_B	● ● ●	● ●	●	●	●
U_D		● ●	●		
U_T	● ● ●				

TABLE 5.1: Bullet points marking the interaction potentials used for each of the explanatory variable sets. In **green**, variable Set 1; in **orange**, variable Set 2; in **blue**, variable Set 3. Each set includes increasingly more water interaction data, with Set 1 containing only ligand-protein interaction data and ligand rotomer data, Set 2 including ligand-holo water data and Set 3 adding apo water interactions.

5.2.2 Functional form of the model

Classically, empirical scoring functions, such as X-Score and Chemscore, have been trained using linear regression, so that $f(U_1, U_2, ..., U_n)$ is linear with respect to each U_i. In recent years, however, there has been a growing trend to use more advanced machine learning techniques[232–237], facilitated by the ever growing amount of structural and binding data[131,136,171,238]. The hope is that such techniques better able at predicting activity cliffs and capturing cooperative protein-ligand interactions that induce relative affinities in ligands that are otherwise inexplicable with a linear model[15,239]. With regards to our water-based scoring function, a non-linear model is necessitated if the contribution an *apo* water makes to a complex's binding free energy depends on its environment prior to displacement *and* whether it is displaced by a polar/non-polar group or not.

In order to test whether increasing model complexity improves the accuracy of scoring functions, we experiment with linear and non-linear regression. As the full model will contain descriptors for protein-ligand as well as water interactions, the regression method

should be robust to a large number of explanatory variables. For linear regression, we use elastic net regularisation as our fitting procedure[240]. This method is a combination of ridge and lasso methods[144], and increases model robustness by limiting the sum of the absolute magnitudes of the regression coefficients (l_1 norm) and sum of the squared regression coefficients (l_2 norm), at the cost of having two free parameters that can be optimised on a validation data set or by using cross validation. The l_1 parameter promotes model sparsity by forcing some regression coefficients to be zero and thus provides an automatic variable selection method. The l_2 parameter, on the other hand, reduces model variance by shrinking the size of the coefficients of correlated descriptors.

For non-linear regression, we use gradient boosted trees (GBTs)[241]. Briefly, GBT regression is a type of ensemble learning technique that fits a succession of weakly performing regression trees in a stepwise manner. When fitting, each new tree is fitted to the residuals of the previous trees. GBTs are a type of additive model and their complexity is primarily controlled by two free parameters. The first is the total number of trees in the model, which is formally similar to the l_1 parameter of the elastic net[144], as simpler models can be selected for by using fewer trees (as opposed to explanatory variables in the elastic net case). The second is the depth of each tree, which determines the maximum number of non-linear interactions between the explanatory variables. Finally, like elastic net regularisation, tree based models are robust when used with a large number of explanatory variables, as each variable must compete with the others to be used to create a split in a tree[144].

5.2.3 Data sets and preparation of structures

We trained and tested various scoring functions on X-ray crystal structures of protein-ligand complexes with experimentally measured binding affinities. We used 10 data sets in total, which were constructed using the 24th September 2010 version of the CSAR

high quality data set of protein-ligand complexes[47] and the 2011 version of the PDBbind data set[28]. The CSAR data set was used to construct one training set (207 complexes) and our validation test set (83 complexes) that both consist of a diverse range of protein-ligand complexes. The 2010 CSAR data set is provided with a list that contains all of the complexes assembled into groups of 90% protein sequence identity. One complex from each group was selected at random to form the training set. The complexes that did not share 90% sequence identity with any other protein were also selected for the training set. From the complexes that remained, another random selection was made to form the diverse validation set, so it too was composed of proteins with less than 90% sequence identity. This validation set is labelled as data set A in Table 1 and in the Results section. As discussed, both the linear (regularised by the elastic net) and non-linear (gradient boosted tree) regression methods have two free parameters. These free parameters were chosen as the pairs that gave the lowest mean absolute error on data set A.

Data set	Label	N^o of Complexes
Diverse training set	-	207
Diverse test set	A	83
HIV-1 protease	B	108
Trypsin	C	66
Factor Xa	D	43
Carbonic Anhydrase II	E	40
PTP*	F	38
Thrombin	G	36
OppA	H	32
Urokinase	I	31

TABLE 5.2: The data sets used for training and testing of the scoring functions in order of data set size. *PTP protein tyrosine phosphatase.

The PDBbind data set was used to assemble 8 single-protein test sets which are labeled as B-I. In 2009 Cheng et al. carried out a comparative assessment of thirty-three popular scoring functions on data sets consisting of HIV-1 protease, trypsin, carbonic anhydrase II and thrombin, using the 2009 edition of the PDBbind database[28]. Although our corresponding data sets have selected from the updated 2011 edition of the PDBbind

data set, we use complexes from the 2009 edition when comparing the accuracy of our models to those in Cheng et al.'s assessment. Factor Xa (C) and OppA (H) have been selected due to the reported importance that apo and holo waters have with respect to the affinities and binding modes of their corresponding ligands[124,242]. Thus, one may expect an explicit water scoring function to have a higher accuracy on these complexes compared to a "dry" scoring function. Finally, PTP (F) and urokinase (I) have been selected due to their pharmacological interest and drug-like ligands. No single complex appeared in more than one data set.

Crystal structures were protonated and prepared for scoring using AutoDockTools[177]. Using the WaterDock method, water molecules were predicted in the presence and absence of the ligand for each protein-ligand complex using the method described in Chapter 3. For large ligands, the docking search volume was extended so that there was at least 2 Å from the edge of box to the ligand. Each of the interaction terms outlined in the Section 5.2.1 were applied using specially written software in python. Regression and analysis was carried out using the statistical software R[179]. Each of the explanatory variables were standardised to have a mean of zero and standard deviation of one. This step is necessary for fitting with elastic net regularisation.

In order to evaluate the performance of our scoring functions in a virtual screening context, two single-protein data sets were taken from the enhanced dictionary of useful decoys (DUD-E)[32]: glutamate receptor ionotropic, AMPA 2 (GRIA2) and HIV-1 integrase. DUD-E provides a single crystallographic structure of each protein, which has been preselected to be most amenable to docking. While DUD-E contains decoy ligands that have selected to have properties matching active ligands and are presumed not to bind, we only tested our scoring functions against experimentally verified inactive compounds. These proteins were chosen for their substantial number of experimentally

tested compounds and their pharmaceutical interest. The GRIA2 data set had 297 active compounds and 201 inactive compounds, each measured to have IC_{50} less than 100 μM. The HIV-1 integrase had 211 active compounds, and 268 inactive compounds, each measured to have IC_{50} less than 50 muM. AutoDock Vina was used to dock each ligand and the top pose was re-scored with our own scoring functions.

5.3 Results

The linear model with only protein-ligand interactions (explanatory variable Set 1) compares favourably to the state of the art in scoring accuracy. Table 5.3 shows the ranking ability of this scoring function on the four single-protein data sets used by Cheng et al. to test the accuracy of thirty-three scoring functions[28]. Our simple linear model outperformed all models on the thrombin data set and was amongst the top performing scoring function for Trypsin and HIV-1 protease, although ranking power on latter data set is low. Our model has a lower than average ranking power on carbonic anhydrase II, although on this data set the well established scoring functions GlideScore XP and GoldScore performed significantly worse than ours, with Spearman rank coefficients of 0.10 and 0.08 respectively. Unfortunately, the study by Cheng et al. did not report the absolute errors of the scoring functions, only the standard deviation from a linear fit. To test our model further, we compared its error and ranking power to AutoDock Vina's scoring functions on all the data sets. Figure 5.2 shows that out of the eight single-protein test sets, six have a lower error and five have more ligands correctly rank ordered compared to AutoDock Vina's scoring function. It has been observed that the number of heavy atoms in a ligand correlates with affinity[231] and can have a better ranking ability than more complicated affinity models[243]. Our simple linear model has a better ranking ability on all of the data sets, with the exception of PTP (G) and urokinase (I). These results suggest that our simplest linear model is representative of

a "good" scoring function, and an adequate control for the addition of water detail and greater complexity.

HIV-1 protease

Scoring function	ρ^a	R^b	SD^c	Better than^d
Our model	0.29	0.29	1.6	79%
X-Score::HPScore^e	0.34	0.34	1.5	
DS::LUDI1^f	0.01	0.01	1.6	

Trypsin

Scoring function	ρ^a	R^b	SD^c	Better than^d
Our model	0.76	0.70	1.2	61%
X-Score::HSScore^e	0.82	0.82	1.0	
GOLD::GoldScore^f	0.05	0.15	1.8	

Carbonic anhydrase II

Scoring function	ρ^a	R^b	SD^c	Better than^d
Our model	0.41	0.65	1.0	39%
DS::PLP2^e	0.77	0.80	0.8	
GOLD::GoldScore^f	0.08	0.36	1.3	

Thrombin

Scoring function	ρ^a	R^b	SD^c	Better than^d
Our model	0.76	0.76	1.3	100%
DS::PLP1^e	0.67	0.69	1.5	
DS::PMF04^f	0.02	0.12	1.9	

TABLE 5.3: Comparison of our scoring function to most and least accurate scoring functions (out of thirty three) tested in the study by Cheng et al.[28]. ^{a} Spearman rank coefficient. ^{b} Pearson correlation coefficient. ^{d} Standard deviation in linear correlation (in $\log K_d$ units). ^{d} The fraction out of thirty three (in percent) that our scoring function has a higher Spearman rank coefficient than. The scoring functions with the highest ^{e} and lowest ^{f} ranking power in Cheng et al.'s study. These results show our scoring function is representative of a top performing function.

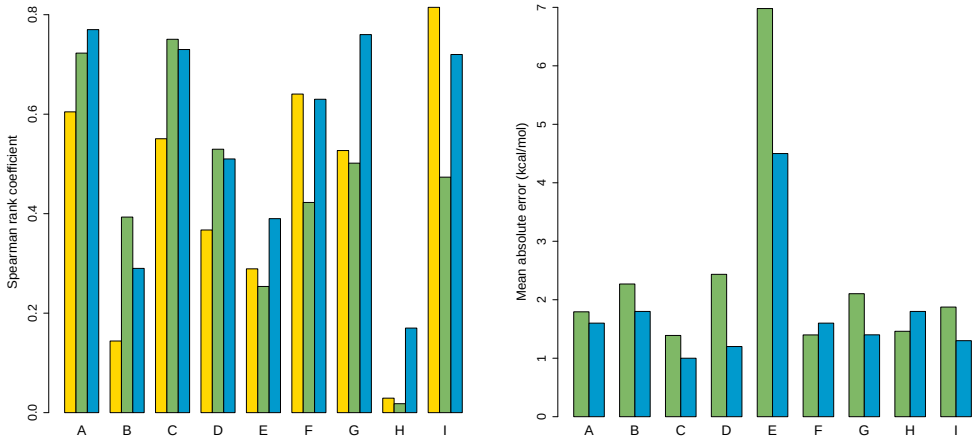


FIGURE 5.2: Comparison of our simple linear model (blue) to AutoDock Vina's scoring function (green) and number of heavy atoms in the ligand (yellow) on data sets A to I. Our scoring function has a higher ranking ability and a lower error on average than AutoDock Vina's scoring function.

In total, six scoring functions were developed. Three linear models and three non-linear models, both trained with three sets of explanatory variables which incrementally increased the amount of explicit water detail. Figure 5.3 shows how the Spearman's rank coefficient and mean absolute error of these models compare. Importantly, there is no overall "best" model. However, the linear models are more robust than the non-linear models, particularly with respect to ranking power on the carbonic anhydrase II, PTP and urokinase data sets, where the non-linear model has no ranking power when all the explanatory variables are used (Set 3). The non-linear model has the largest improvement from the simplest linear model on HIV-1 protease when using explanatory variable Set 2, which includes holo-water information. However, in this case, Spearman's rank coefficient is only 0.42. To the best of my knowledge, the only comparison between a linear and non-linear regression techniques using the same explanatory variables was done by Head et al in 1996[244]. By comparing a partial least squares (linear) model to a neural network (non-linear) model, they also found the linear model to be more robust.

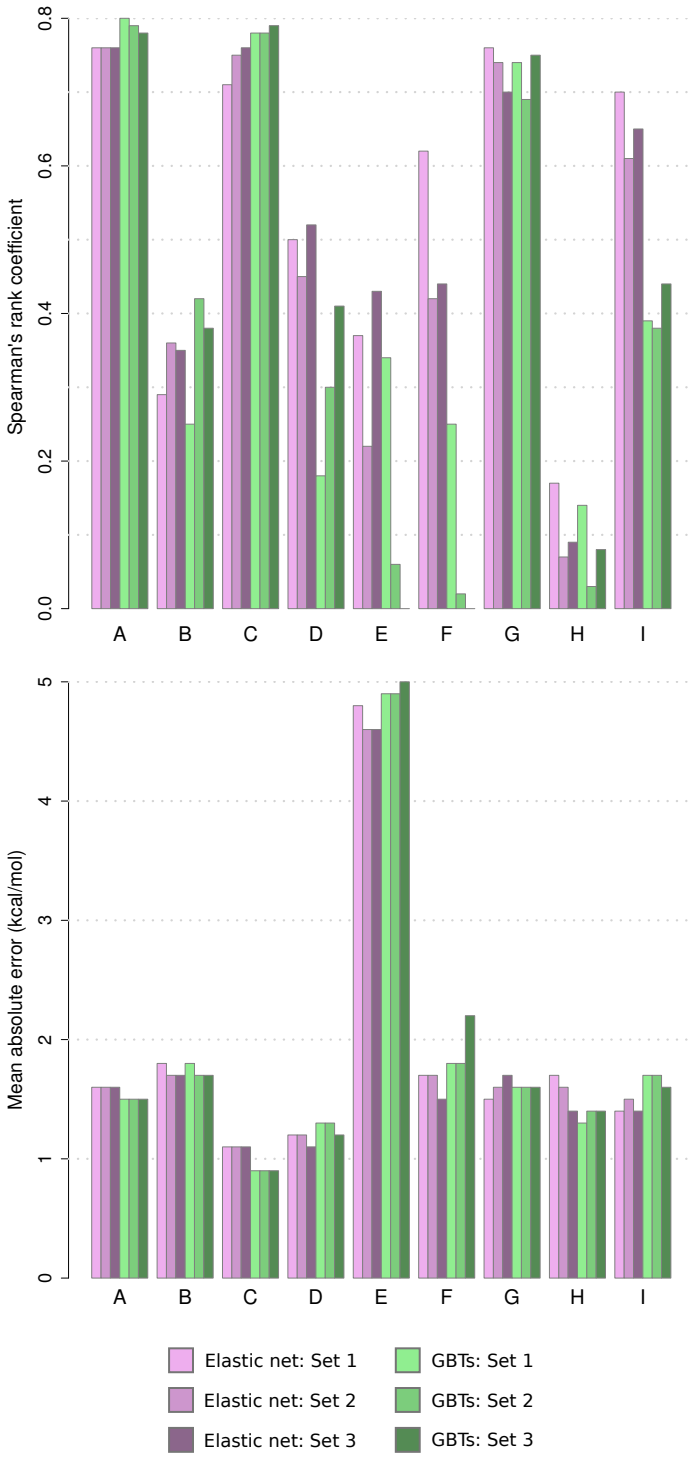


FIGURE 5.3: Bar graphs showing Spearman's rank coefficient and mean absolute error of the different scoring functions when applied to single-protein data sets (B to I) after being trained on a diverse range of proteins. Data set A was used to calibrate the free parameters of the fitting procedures. The scoring functions are composed of 2 different regression methods, linear with elastic net regularization (Elastic Net) and gradient boosted trees (GBTs), each with 3 different sets of explanatory variables. Explanatory variable Sets 1 to 3 contain increasingly more explicit water data. Numerical values of the perfomance measures (including R^2) for the linear and non-linear models, with explanatory variable Set 1, are shown in Tables A.3 and A.4 of the Appendix. Increasing the complexity and the water content of the models does not significantly improve accuracy. Scoring function error is highly test set dependent, irrespective of the number of explanatory variables and regression method.

Inspection of Figure 5.2 reveals the two models that have consistently the highest Spearman's rank coefficients are the simplest linear model (with variable Set 1), with no explicit water detail, and the linear model with all the explicit water interactions terms (variable Set 3). While the linear model with Set 3 performs as well as the linear model with Set 1, it has a significantly lower ranking ability on PTP (F). On this data set, the linear model with variable Set 1 has a Spearman's rank coefficient of 0.62 compared to 0.44 with variable Set 3. The R^2 of the fits are also significantly different on this data set, with values of 0.32 compared to 0.19, using variable Sets 1 and 3 respectively.

As discussed in Section 5.2.3 the training set and validation set (data set A) are composed of protein-ligand complexes with less than 90% sequence similarity. Typically in regression, the model that performs best on the validation set is the one that is selected for further validation or used in a predictive context[144]. If we had chosen our primary scoring function in this way, one of the GBT models would be selected owing to their higher ranking ability and lower errors on data set A (see Figure 5.3). Yet these scoring functions perform significantly worse on data sets D, E, F and I than the linear models. It seems then, that the accuracy over a diverse range of protein-ligand complexes is not indicative of the performance on individual proteins. This has also been observed with the scoring function RF-Score[235], which although reported to be the most accurate on a diverse range of proteins and ligands, was later found to be very inaccurate for individual proteins[245].

Previous studies have found that the enrichment rates of active compounds is highly test set dependent, and scoring functions with high enrichment rates on some proteins can be no better than random on others[25,26,31,33,194]. Inspection of Figure 5.3 reveals a similar trend, as not one of our models has the lowest error and ranking ability on all data sets. As a scoring function's error and ability to rank compounds is correlated with its ability to select active compounds from a set of inactives[214], we would expect a

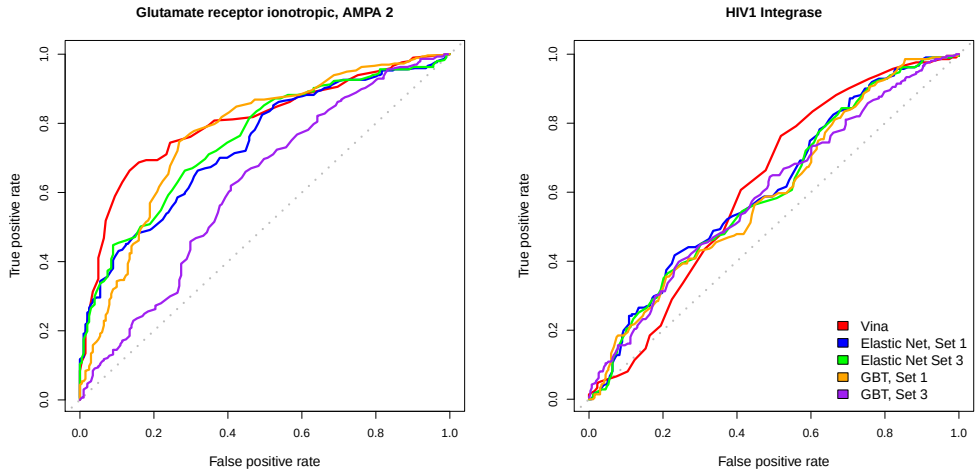


FIGURE 5.4: Receiver-operating characteristic curves for glutamate receptor ionotropic, AMPA 2 and HIV-1 integrase for selecting compounds with IC50s less than 50μM (actives) from a library of compounds also containing compounds with affinities greater than 50μM (inactives).

similar variability of our models on such tests. Nevertheless, is informative to evaluate our scoring functions in a virtual screening context. Thus, we opted to test and compare our scoring functions enrichment capabilities on only two proteins, with the proviso that we do not expect similar performances on other proteins.

Figure 5.4 shows the receiver-operating characteristic (ROC) curves for glutamate receptor ionotropic, AMPA 2 and HIV-1 integrase when our scoring functions are used to re-score ligand binding poses predicted by AutoDock Vina. Table 5.4 shows the area under the curve (AUC) and enrichment factors at 5% and 10% of the top predictions. The AUC is equal to the probability that a scoring function will correctly differentiate between a randomly chosen active compound from a randomly chosen inactive compound. The enrichment factor at a given percent of the data set is arguably a more valuable measure of a virtual screen's success, as in practice, only a small fraction of a vast chemical library will be experimentally assayed.

Inspection of Figure 5.4 and Table 5.4 shows that on both data sets, none of the scoring functions improve on AutoDock Vina's performance, although the accuracies are similar. This is despite the fact that our scoring function has a consistently higher ranking

power on data sets B to I. This discrepancy may simply be due to chance, given the data set dependance of scoring accuracy, or may be due to the way that our scoring function is trained. While ranking ability and enrichment are correlated, enrichment may be improved by specifically training models to distinguish between active and inactive compounds. This was the strategy behind the freely available NNScore, a scoring function composed of an ensemble of artificial neural networks[234]. By docking decoy compounds into proteins and assigning these complexes as having binding energies of zero, the neural networks were trained to distinguish decoy complexes from crystallographic complexes with binding free energies that were experimentally measured to be negative (favourable). The most recent iteration of NNScore claims to be of comparable accuracy to a commercially available virtual screening package[33]. One can also envisage a similar strategy for binding mode prediction, where a classifier is trained to distinguish decoy poses from crystallographically derived poses, although to the best of our knowledge, this has yet to be implemented.

Tellingly, enrichment is *not* improved by the addition of all the water detail on either HIV-1 integrase or glutamate receptor ionotropic, AMPA 2. However, given the variability of our scoring functions on data sets B to I (see Figure 5.3), it may be that by chance, water detail improves enrichment rates on as yet untested proteins.

Glutamate receptor ionotropic, AMPA 2

Scoring function	AUC	EF_5	EF_{10}
AutoDock Vina	0.80	1.68	1.58
Elastic Net: Set 1	0.74	1.68	1.61
Elastic Net: Set 3	0.75	1.68	1.61
GBTs: Set 1	0.78	1.47	1.41
GBTs: Set 3	0.62	1.29	1.23

HIV-1 integrase

Scoring function	AUC	EF_5	EF_{10}
AutoDock Vina	0.62	1.02	0.88
Elastic Net: Set 1	0.61	0.99	1.20
Elastic Net: Set 3	0.60	0.89	1.32
GBTs: Set 1	0.60	1.17	1.45
GBTs: Set 3	0.60	1.56	1.36

TABLE 5.4: Comparison of the area under curves (AUC) and enrichment factors at 5% (EF_5) and 10% (EF_10). In general, the accuracy of our scoring functions are similar to that of AutoDock Vina's. Note that all explicit water detail (Set 3) does not reliably improve the enrichment factors on both data sets compared to only protein-ligand descriptors (Set 1).

It is interesting to note that the ligands in the OppA data set (H) are poorly ranked with all of our models, and that all our scoring functions have a better ranking ability with trypsin (C) and thrombin (G) than factor Xa (D). A motivation for including both OppA and factor Xa as test sets was to test whether a higher accuracy could be achieved on these data with explicit water scoring functions given the importance of water in those systems[124,242]. However, a high correlation of predicted affinities and high ranking ability of a scoring function on these data sets will require a very low average error. Following the definition of a Pearson's and Spearmans's correlation coefficients, the lower the standard deviation of the response variable in a data set, the more accurate

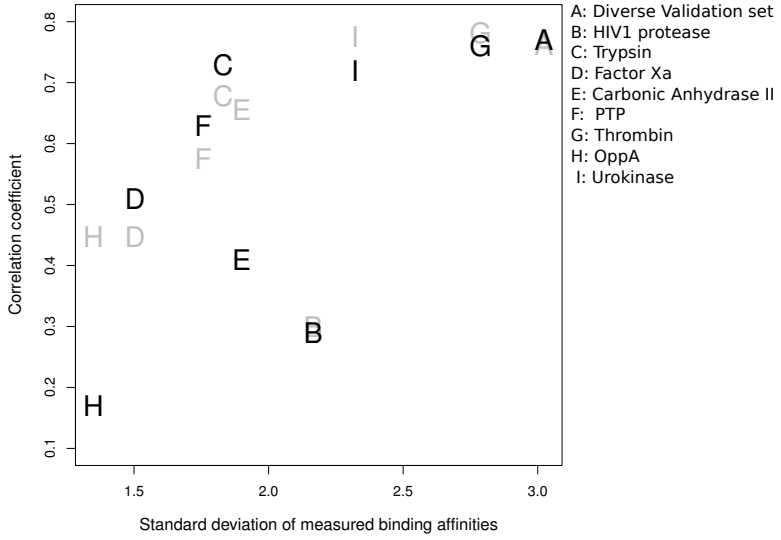


FIGURE 5.5: Standard deviation of the experimental binding affinity in data sets A-I, plotted against Spearman's rank coefficient (black) and Peason's correlation coefficient (grey). This shows that the ability of a scoring function to rank order compounds is strongly determined by the spread of the affinities in a data set. HIV-1 protease (B) and carbonic anhydrase II (E) have a lower correlation than expected given the standard deviation of their affinities.

a model has to be in order to achieve a satisfactory correlation (see Chapter 2, Section 2.3). Figure 5.5 shows the standard deviation of affinities of each data set plotted against Spearman's rank coefficient and Pearson's correlation coefficient. It is clear that achieving a high correlation with ligands bound to OppA and factor Xa will be very difficult owing to their respective standard deviations of 1.3 kcal/mol and 1.5 kcal/mol in the experimentally measured binding affinities.

As a further validation, we compared our linear models with explanatory variables Sets 1 and 3 to results obtained using two early versions of WaterMap[124]. WaterMap is a semi-empirical model that uses the computed entropies and enthalpies of apo hydration sites that are occupied by the ligand. Hydration site locations and thermodynamics are predicted using short molecular dynamics simulation. While primarily designed to predict relative binding free energies of congeneric ligands, its ability to predict absolute binding free energies was also investigated on twenty-eight ligands complexed with factor Xa (from a different data set to our data set D). They experimented with two models

that had three and five free parameters each. The root mean squared error (RMSE) and R^2 of these models after leave-one-out cross validation are shown in Table 5.5. Using the same complexes and the same experimental affinities they reported, both of our models substantially outperform the WaterMap predictions in both error and correlation. This comparison, in addition to our own water model analysis, suggests the addition of explicit water detail is not necessary to improve scoring accuracy. While there have been further modifications to the WaterMap method since the factor Xa study[130], metrics summarising the quality of the predictions were not reported.

	Linear Model		WaterMap	
Metric	Set 1	Set 3	(a)	(b)
RMSE (kcal/mol)	1.2	1.6	3.0	1.8
R^2	0.40	0.41	0.11	0.31

TABLE 5.5: Comparison between the linear models with explanatory variable Sets 1 and 3 to early versions of WaterMap[124]. WaterMap version (a) had three fitting parameters and version (b) had five, and the errors reported are from leave-one-out corss validation. The root mean squared error is denoted as RMSE. Both of our linear models have a lower error and higher correlation than either versions (a) and (b) of WaterMap.

5.3.1 Water importance in affinity predictions

Why did the full water-based scoring function not improve the accuracy of the simplest protein-ligand interaction model? The largest set of explanatory variables (Set 3) includes a total of 20 interaction terms that are designed to capture contributions to protein-ligand binding affinity, such as the shape complementarity between the protein and ligand and the displacement of apo water molecules. One of the benefits of using elastic net regularisation and gradient boosted trees (GBTs) for regression is that both methods can be used to find out which interaction terms are "important" with regards

to affinity prediction. An importance analysis allows us to gain insight into which of the explicit water interactions are the most relevant for rapid affinity predictions.

We denote N explanatory variables as $X = (X_1, X_2, ..., X_N)$ and a scoring function as $f(x)$. A measure of the relative influence an explanatory variable X_i has on a scoring function $f(x)$, as proposed by Friedman[241] is

$$J_i = \sigma_i \sqrt{E\left[f_i^2(x)\right]}, \tag{5.9}$$

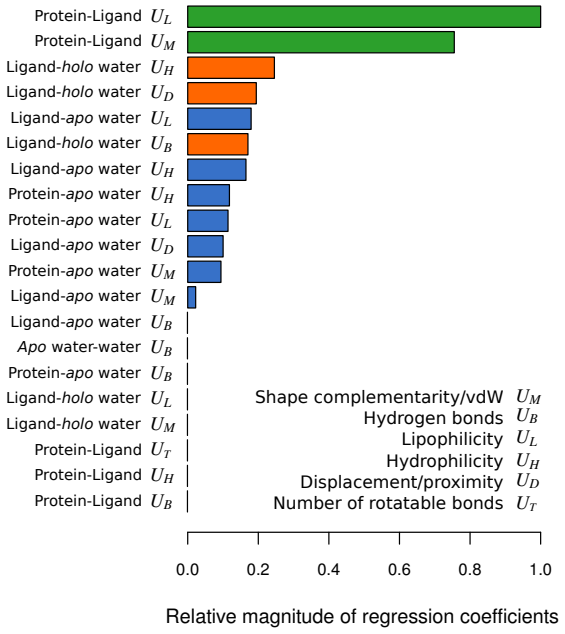
where σ_i is the standard deviation of variable X_i, $f_i(x)$ is the partial derivative of the scoring function with respect to variable X_i, and the expectation, denoted by E, is the average over all X. Calculating J_i for each variable is equivalent to performing a type of sensitivity analysis[246]. Prior to performing regression, we standardised the variables to have a standard deviation equal to one. Hence, for our linear models, J_i reduces to the magnitude of the regression coefficient of X_i. For our GBT models, it is not possible to evaluate J_i directly, as partial derivatives cannot be evaluated over the discrete splits in the regression trees. Instead, following Breiman and Friedman[241], we approximate Equation 5.9 by the mean reduction in the error of a model that occurs when X_i is used for a new split in the tree, averaged over all trees. Figure 5.6 shows these importance measures evaluated on the linear and GBT models when fitting with all the variables (Set 3). The free parameters of the fits were optimised on the diverse protein-ligand validation set (data set A) as before.

Figure 5.6 shows that despite the difference in functional forms, the three most important terms are the same for the elastic net and GBTs models. The interactions with by far the largest influence on both the linear and non-linear models are the vdW force/shape complementarity between the ligand and protein and the degree to which lipophilic

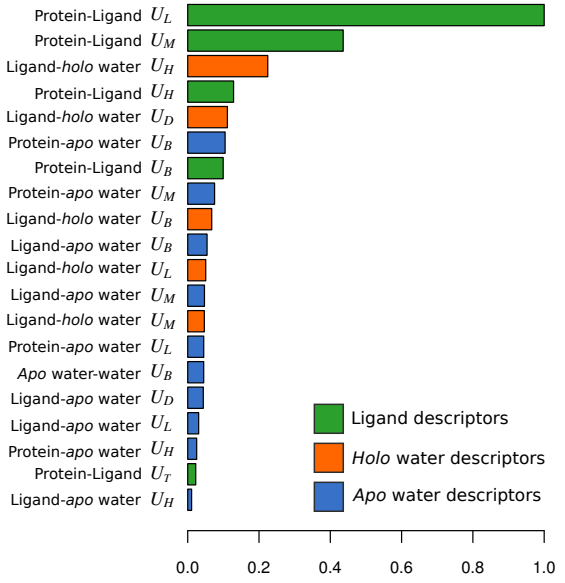
ligand atoms are buried in apolar cavities. As most of the binding affinity information is contained in these protein-ligand based terms, it is perhaps not surprising that the accuracy of the linear model with explanatory variable Set 3 is largely the same as the simplest protein-ligand interaction model. In both models, the third most important term, with less than 30% of the influence as the protein-ligand lipophilic term, is the ligand-holo water hydrophilicity interaction. This term operates much like a hydrogen bonding term, and is high whenever a holo water is near a polar ligand atom. This suggests the largest explicit water effect most easily captured by an empirical model is the ability holo waters have to bridge polar interactions with ligand, and not apo water displacement. It is interesting to note that the number of rotatable bonds in the ligand has very little influence in the GBTs model and zero influence in the linear model. It was included to try to capture the entropic penalty of freezing a ligand in a particular binding mode. Despite being a very popular term in empirical scoring functions[217,232,234,244], this term is unnecessary, and without using a regularisation technique as we have done with the elastic net, its inclusion is likely to have only added noise to those previous models.

5.3.2 Water placement error

Critical to the preceding work is how trustworthy our results are, and whether we can expect them to generalise to other rapid affinity methods that use different water placement and scoring methods. If we had incorporated explicit water differently in our scoring functions, could we expect to see a similar results? Whilst we have shown in Chapter 3 that the WaterDock method is largely accurate in predicting hydration sites, it is based on the stochastic docking program AutoDock Vina. Even though three docking runs are averaged over to produce a prediction, stochasticity still remains in the predictions. We can understand how this stochasticity will affect the error of a scoring function by the bias-variance decomposition of an empirical model's error, which was discussed in



(a)



(b)

FIGURE 5.6: Relative sensitivity of the (a) linear and (b) GBTs models to the explanatory variables in Set 3, which includes all water detail. Even though GBTs are non-linear, the ordering of the three most important terms is the same as the linear model. In both models, the protein-ligand lipophilic interaction and vdW interaction dominate the explicit water-based terms, which explains why the addition of explicit water detail does not improve scoring function accuracy. In (a): as the linear model was trained using elastic net regularisation, some of the regression coefficients are automatically set to zero during the fitting process.

Chapter 2, Section 2.3. The mean squared error of a model can be decomposed into a sum of its intrinsic error, its squared bias and the variance of model due to re-sampling of data. Any stochasticity in the explanatory variables will increase the *variance* of the scoring function. In the following equation, the uncertainties in the water molecule locations are related to the variance, and therefore the error, of a scoring function. To do this in an easily comprehensible way, it is assumed that the explanatory variables are normally distributed, although generalisations to other distributions exist. The variance of the scoring function is denoted $\mathrm{var}\big[f(x)\big]$, and from Cacoullos and Papathanasio[247], we have

$$\sum_{i=1}^{N} \sigma_i^2 E^2\big[f_i(x)\big] \leq \mathrm{var}\big[f(x)\big] \sum_{i=1}^{N} \leq \sigma_i^2 E\big[f_i^2(x)\big]. \tag{5.10}$$

Equation (5.10) shows that the larger the uncertainty regarding an explanatory water variable, represented through σ_i^2, the larger the variance of the scoring function. A large variance implies a scoring function has a large error. Comparing Equation 5.9 to 5.10, one can see that the square of Friedman's importance measure appears in the upper bound for the variance of a scoring function. In addition, the appearance of partial derivatives of the scoring function in the lower and upper bounds of $\mathrm{var}\big[f(x)\big]$ highlights the close relationship between a model's sensitivity and its variance.

We investigated the impact of the uncertainty in water molecule locations on our scoring functions and data sets. Using the data sets shown in Table 5.2, the location of the apo and holo water molecules were predicted twice. The first set of locations were used for all of the preceding scoring function training and testing. The second set is used to test the consistency of the predictions. On average, 7.7% of the first set of the predictions were not found within 1 Å of the second set. When the linear model with

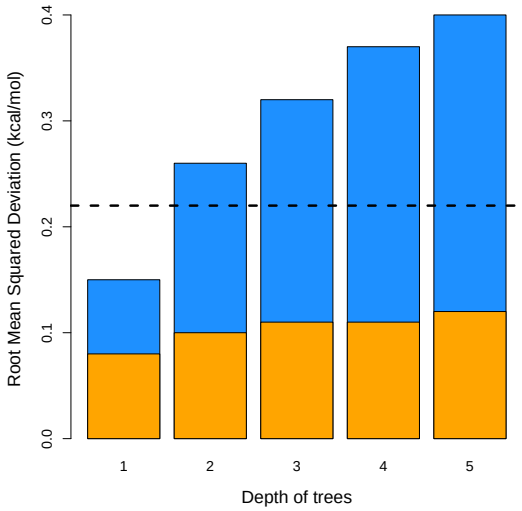


FIGURE 5.7: Barplot showing how increasing the non-linearity of a scoring function amplifies the error due to the uncertainty in the water positions. Gradient boosted trees (GBTs) were used for regression. The depth of each tree sets the maximum number of non-linear interactions in the models. Blue bars show the root mean squared deviation (RMSD) of predictions on the training set when using two sets of apo and holo water locations. As GBTs have a stochastic fitting element, orange bars show the RMSD of a second GBT model on the same water positions. The dashed line shows the RMSD of the linear water model on the two sets of water positions. Above a depth of two, GBTs have a larger error due to water uncertainty than the linear model.

explanatory variable Set 3 is re-applied to the training set but with the second set of water locations, the root mean squared deviation (RMSD) of the predictions was found to be 0.18 kcal/mol. Similarly, the average RMSD of the predictions using the two water sets on the single-protein data sets (B-I) was 0.29 kcal/mol. From the bias-variance decomposition of the error, these RMSDs contribute to the error of the water model. We found that averaging over the two sets of predictions did not lower the mean error or ranking performance of the scoring functions. The first and second set of predicted water locations that are within 1 Å were used to test whether using only these "consistent" sites improved accuracy. Re-training and testing the linear model with variable Set 3 with these consistent sites yielded lower (worse) Spearman's rank coefficients by at least 0.10 on factor Xa, carbonic anhydrase II and urokinase. The ranking ability on the other complexes remained largely unaffected. This suggests that removing the less consistent water molecules lost information relevant to affinity prediction.

A motivation of using GBTs was to be able capture any non-linear cooperative effects due to protein, ligand and water interactions. A free parameter of GBTs is the depth of each tree, which determines the maximum number of non-linear interactions in a model. Figure 5.7 shows how increasing the depth of each tree in the GBT steadily increases the RMSD of the predictions on the training set when using the second set of water molecule locations. Thus, while it may be that non-linear interactions between the explicit water variables and protein-ligand terms are important for accurate affinity calculations, accounting for such effects is very difficult if there is uncertainty regarding the values of the explanatory variables. As Figure 5.7 demonstrates, uncertainties regarding explanatory variables will increase a scoring function's error as the non-linearity is also increased. In the case of explicit waters in scoring functions, this suggests a deterministic method, such as GRID[162], may be better suited for scoring with explicit waters. However, it is important to note that stochasticity and uncertainty form an integral part of modern ligand docking techniques. Finding the best scored binding mode of a ligand requires a high dimensional search over conformation space. To speed up this search, most popular docking programs utilise types of stochastic algorithms[248], such that repeated dockings may yield slightly different ligand poses. Through Equation 5.10, uncertainty in a binding mode can increase the expected error of a scoring function. As our analysis has used crystallographic binding modes, our reported errors are likely to be much better than what can be achieved when using docked poses. Owing to the greater sensitivity of our models to protein-ligand interactions than to explicit water interactions (see Figure 5.6), we expect the error due to binding mode uncertainty to be much larger than our reported error due to water position uncertainty.

5.4 Summary

We developed our own scoring function which compares favourably to other popular scoring functions. Apo and holo water molecules, as predicted by WaterDock, were used to add terms to our scoring function to account for bridging water interactions and apo water displacement by a polar and apolar ligand groups. This explicit water detail did not improve the accuracy of our original scoring model and the protein dependent nature of scoring accuracy was left unchanged. Through a sensitivity analysis of linear and non-linear water models, we determined this was because the greatest contributions to predicted binding affinity are due to protein-ligand interactions and not our explicit water descriptors – the addition of which, increased the variance and hence the error of the scoring functions.

Chapter 6

The limits of empirical scoring functions

6.1 Introduction

Despite the continual development of new scoring functions, comparisons of different scoring functions by affinity prediction accuracy[28,245] and virtual screening enrichment[25,26,30–33] have consistently demonstrated that performance is heavily dependent on the protein studied. The scoring functions developed in Chapter 5 were similarly variable, despite that fact that – in contrast to other scoring functions – the models incorporated explicit water information. While the maximal ranking ability of a scoring function is limited by the standard deviation of the test set affinities, the reason why different scoring functions can have wildly different accuracies on the same data set has not been previously investigated. This difficulty, also being observed with different paramerisations of GBSA and PBSA models[38,249,250], led Page et. al.[249] to remark that

> "...the fact that no method is consistently better highlights the inherent Catch-22: it is only possible to determine the method of choice for a given class of problem if the answer is already known."

This comment highlights an opinion in the field that there is no "one size fits all" solution, and that each method should be investigated and applied on a case-by-case basis. In contrast, the goal for many remains the development of a fast universally applicable model[208]. This aim is implicitly invoked whenever a scoring function is trained and tested on a wide varieties of proteins and a ligands. A number of studies have sought to improve generalised scoring function accuracy by employing advanced machine learning techniques, many protein-ligand interaction descriptors and large training sets[232–237,251]. The hypotheses underlying each of these studies are neatly summarised by Zilean et.al.[251]

> "Improvements of empirical scoring functions can be envisaged along three lines: first, by developing new descriptors; second by training on larger high-quality data sets of protein-ligand structures with experimentally determined affinities; and third, by using alternative methods for regression analysis."

Our investigations in Chapter 5 also followed similar lines of enquiry, as we experimented with a non-linear machine learning regression method and increased the number of descriptors to account for water effects. While non-linearity and larger descriptor sets improved accuracy in some cases, our results agreed with that of Kramer's[245] and Durrant's[33], as scoring accuracy remained test set dependent, despite the increase in complexity of the model. It has also been reported that scoring functions that have been optimised specifically for one protein are more accurate than generalised models[252,253]. The sheer diversity of scoring functions that have little transferability suggests that there is an underlying cause which is common to all.

Given the potential impact on drug discovery and the substantial effort in scoring function development, it is vital to understand the fundamental uncertainties in these rapid affinity models. There remain very few analytical studies on sources of error in affinity prediction methods. Notably, Faver et. al. investigated the systematic and random

errors associated with the interaction energies of protein-ligand complexes[254]. Using a fragment based approach[255], they reasoned that the random error in electrostatic interaction calculations rises with the system size. Yet, questions still remain with regards to the limits of scoring function accuracy. For instance, is it possible for a particular set of descriptors and functional form of a model to achieve a negligible error? Also, can a universally applicable scoring function ever be better than a targeted one? Without a formal analysis of the structure-based modeling process, questions such as these cannot be answered fully.

In this chapter, we investigate the inherent uncertainties in empirical structure-based models with a rigorous mathematical analysis. In particular, we focus on the factors that limit the transferability of empirical models. We utilise statistical learning theory and information theory to ensure that our analysis is independent of the method of regression, the size of the training set and the scoring function descriptors. We verify our theoretical predictions with our own scoring functions from Chapter 5.

6.2 Protein-ligand structures have unique probability distributions

Chapter 2, Section 2.3 discussed that when performing regression, such as when fitting a scoring function to experimental data, one implicitly utilises statistical learning theory. This theory treats a data set – in our case a set of protein-ligand structures and their corresponding affinities – as though it has been generated by a probabilistic process. A successfully trained model is one that has captured the true, relevant statistical properties underlying the data set. For the model to then make accurate predictions, any future data must be generated by the same probabilistic process as that which generated the training set.

We model data sets of protein-ligand complexes and their corresponding affinities as though they have been sampled from a probability distribution. We denote affinity as Y and structural descriptors as X. To be as general as possible, we will not state what X is: it may represent a single variable, such as interaction energy, or a vector of empirical descriptors. The convention in statistics is that capital letters denote random variables, while lower case letters denote realisations of the variable and are used in functions. This convention will not be followed when conflicting with common notation in physics.

As Equation 2.14 in Chapter 2 shows, the function that minimises the mean squared error (MSE) is given by the conditional mean of Y for a given X. Thus, in this section, we ascertain the statistical relationship between X and Y to begin to understand some fundamental causes of scoring function regression error.

In rigorous free energy calculations[40,41], a control parameter, denoted λ, is often used to define a molecular system or set of constraints on that system. In ligand binding free energy calculations, the control parameter associated with the bound state, denoted λ_b, is switched over the course of possibly many simulation windows to the value associated with the unbound state, denoted λ_u. As free energy is a state function, binding affinity Y depends only on λ_b and λ_u. Experimental error gives rise to uncertainty in the value of the "true" free energy difference between the two states. We represent this intrinsic uncertainty in the free energy between the states via the probability distribution function (PDF) $s(y|\lambda_b, \lambda_u)$.

In contrast to rigorous methods, scoring functions use a single snapshot of the protein-ligand complex to predict affinity. As sampled from the bound state, any such snapshot is dependent only on λ_b. In Chapter 2, $\mathbf{r}^N$ was used to denote the three dimensional coordinates of an N atom system, which is different for each protein-ligand complex. If the positions of the N atoms in a complex are sampled from the equilibrium distribution of the complex in solvent, from statistical mechanics, the PDF of a structure given λ_b

is simply the Boltzmann distribution from Equation 2.3 integrated over all momenta, given by $p_B(\mathbf{r}^N|\lambda_b)$. As we are considering many different protein-ligand complexes, we denote R as the structural phase-space spanning many protein-ligand complexes, so that the PDF of observing a particular structure given a particular bound state is $p(r|\lambda_b)$.

To understand scoring function error, we consider the probabilistic relationship between structure R and affinity Y. For a given protein-ligand complex, denoted θ, the PDF of observing r and y is given by $p(r, y|\theta)$. The question "what is the binding affinity of a particular protein-ligand complex θ?" is implicitly asking "what is the free energy difference between states defined by λ_b and λ_u?". In other words, θ, is really a surrogate label for λ_b and λ_u. From the above discussion, R and Y are dependent on λ_b and λ_u and not on each other. Thus, R and Y are conditionally independent for a given complex, so that

$$\begin{aligned} p(r, y|\theta) &= p(r, y|\lambda_b, \lambda_u) \\ &= p(r|\lambda_b)s(y|\lambda_b, \lambda_u), \end{aligned} \tag{6.1}$$

Once a particular structure has been sampled, the coordinates of the complex R are processed to yield the structural descriptors that will be used in the scoring function. For universal scoring functions, these descriptors depend only on the structure, which in turn depends on which complex is selected. The process of measuring affinity and obtaining descriptors is therefore a Markov chain, and is shown in Figure 6.1. If the process was not Markovian, the way the structure was sampled would depend on the affinity, or, the structural descriptors used in the scoring function would depend on the complex. The latter does not occur for universal scoring functions. Consideration of

this Markov relationship and Equation 6.1 reveals the PDF for particular structural descriptors and affinity

$$\begin{aligned} p(x, y|\theta) &= \int p(x, r, y|\lambda_b, \lambda_u)\ \mathrm{d}r \\ &= \int p(x|r)p(r|\lambda_b)s(y|\lambda_b, \lambda_u)\ \mathrm{d}r \\ &= \left[\int p(x|r)p(r|\lambda_b)\ \mathrm{d}r\right] s(y|\lambda_b, \lambda_u) \\ &= p(x|\lambda_b)s(y|\lambda_b, \lambda_u). \end{aligned} \tag{6.2}$$

The most important feature of the above is that $p(x|\lambda_b)$ depends on the Boltzmann average of $p(x|r)$ over the position space of the complex θ. By virtue of the fact that the equilibrium statistics of different complexes are not the same, it evident that each complex has a unique Boltzmann distribution. This suggests that each $p(x|\lambda_b)$ may be different for every bound state, in which case $p(x, y|\theta)$ would be unique for each complex. In which case, the assumption in statistical learning theory that the training and test set must be sampled from the same probability distribution would be violated if one transfers a structure-based model from one protein-ligand complex to another. In the next section, we discuss our information theoretic approach that quantifies the error of a missapplied model. This framework shows that using the most accurate set of descriptors are the structures themselves, implying that $p(x|\lambda_b)$ is indeed distinct for each complex.

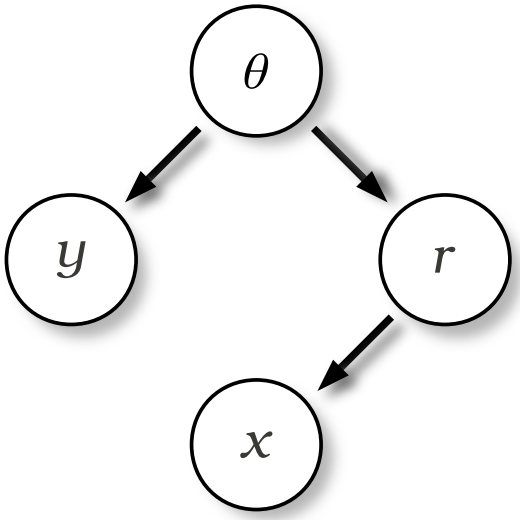


FIGURE 6.1: A graphical model of the data generating process for universal structure-based scoring functions. A given selection of a protein-ligand complex (θ) determines the measured binding energy (y) and observed structure sampled (r). The structure is then processed to create the structural descriptors (x) used in a scoring function. The affinity and structure are conditionally independent given θ.

6.3 Information theoretic approach to scoring function error

Many factors can cloud the reasons for an empirical model's inaccuracy. In structure-based affinity models, one has to choose the representation of a complex – a possible description might be the surface complementarity of a between a drug and a protein – and the functional form of the model. Both choices may introduce errors that are difficult to disentangle from the fundamental uncertainties arising from rapid affinity prediction. We require a method of analysis that is independent of the representation of the complex, the functional form of the model, as well as the method of regression. In this section, we describe our information theoretic approach to scoring function error. This approach is our own, and its construction was a significant undertaking. It has been influenced by Principe[256], who used the entropy of the error distribution as a loss function, and Burnham and Anderson[147], who discuss relative entropy as a measure of how far a model is from the "true" process.

Equation 2.14 showed that for a data set sampled from $p(x, y)$, the functional relationship between structure X and affinity Y which minimises the mean squared error (MSE) is given by the affinity averaged over the conditional PDF $p(y|x)$. What is the error of a model with the lowest MSE on data sampled from a another distribution, $q(x, y)$, if it is applied to data sampled from $p(x, y)$? We assert, and show below, that the expected error of such a model is encapsulated by the cross entropy

$$C(Y|X) = -\iint p(x, y)\, \ln q(y|x)\, \mathrm{d}x\, \mathrm{d}y. \tag{6.3}$$

In a well known information theoretic result[145], cross entropy expands to

$$C(Y|X) = h(Y|X) + D\big(p(y|x)||q(y|x)\big), \tag{6.4}$$

where $h(Y|X)$ is the conditional entropy of $p(x, y)$, and $D\big(p(y|x)||q(y|x)\big)$ is the relative entropy between the two conditional distributions. Equation 6.4 can be understood as representing the minimum uncertainty, or error, of the binding affinity given the structure, plus the uncertainty due to the model; choosing the wrong model only increases the uncertainty of the system. In our information theoretic perspective, $h(Y|X)$ is the minimum achievable error, which is irreducible for a given set of descriptors, and $D\big(p(y|x)||q(y|x)\big)$ is the bias for assuming the structure-affinity relationship is encoded in $q(y|x)$ when in reality it is encoded in $p(y|x)$. This bias is a type of random error, as opposed to a systematic error. Any bias ensures that the minimum error is not achieved for the wrong model, and borrowing a term from financial decision theory[257] and signal processing[258], we shall refer to the positive deviation from the minimum error as *regret*.

Comparison of Equation 6.4 to Equation 2.15 shows that the former is similar to the bias-variance decomposition of the MSE, with $h(Y|X)$ representing irreducible error and $D\big(p(y|x)||q(y|x)\big)$ the bias. There is no variance term in Equation 6.4 as we are considering the probability distributions themselves and not the data that has been sampled from them.

The cross entropy naturally accommodates the errors which occur when transferring models between different probability distributions. This is relevant for scoring functions, as Section 6.2 demonstrated that structure-affinity probability distributions are distinct for each complex. Chapter 2, Section 2.3 discussed how the aim of regression is to find the optimal which is encoded in the conditional PDF of Y given X. If one trains a scoring function on data sampled from $q(x, y)$ and applies it to data sampled from $p(x, y)$, the deviation the optimal model fitted to $q(x, y)$ has from the minimum error is given by $D\big(p(y|x)||q(y|x)\big)$. Focusing on optimal models encoded in conditional PDFs ensures that our analysis is independent of the method of regression and size of the data set.

6.3.1 The most accurate descriptors for structure-based scoring functions

What are the structural descriptors for a structure-based scoring function that give the lowest error? Chapter 2, Section 2.4, discussed how the explanatory variables which maximise the mutual information with the response variable $I(X; Y)$, minimise the conditional entropy $h(Y|X)$: the irreducible error for a given set of descriptors. As discussed and shown in Figure 6.1, selecting a complex; obtaining an affinity measurement and a structural snapshot; and deriving a set of descriptors, forms a Markov chain. This allows us to exploit the Data Processing Inequality (proven and discussed in Chapter 1 of Cover and Thomas[145]). This inequality states that the processing of data, such

as processing structures into scoring function descriptors, cannot increase information. Applying the inequality to the Markov chain in Figure 6.1, one finds

$$I(\theta; Y) \geq I(R; Y) \geq I(X; Y). \tag{6.5}$$

Simply put, the unprocessed structures share *at least* as much information with affinity as the scoring function descriptors share with affinity. Equation 6.5 implies that $h(Y|R) \leq h(Y|X)$, so that the choice of descriptors that has the lowest minimum error are the structures themselves. Thus, we focus on an idealisation of a structure-based model, the error of which is a lower bound of what can be achieved in practice.

6.3.2 Cross entropy and normally distributed errors

To show more explicitly how $C(Y|X)$ is related to the MSE of a model, consider the case when the data sampled from $q(x, y)$ is modeled as $y = f_q(x) + \epsilon_q$, where ϵ_q is white noise with a mean equal to zero with standard deviation σ_q. If the white noise is normally distributed, and X represents a single explanatory variable, then

$$q(y|x) = \sqrt{\frac{1}{2\pi\sigma_q^2}} \exp\left(\frac{-(y - f_q(x))^2}{2\sigma_q^2}\right). \tag{6.6}$$

Inserting $q(y|x)$ into Equation 6.3 gives:

$$C(Y|X) = \frac{1}{2\sigma_q^2} \iint p(x,y)(y - f_q(x))^2 \, \mathrm{d}x \, \mathrm{d}y + \frac{1}{2} \ln 2\pi\sigma_q^2.$$
$$= \frac{1}{2\sigma_q^2} \mathrm{MSE}_{q,p} + \frac{1}{2} \ln 2\pi\sigma_q^2, \tag{6.7}$$

where $\mathrm{MSE}_{q,p}$ is the MSE of the model from $q(y|x)$ applied to data from $p(x,y)$. As $\mathrm{MSE}_{q,p}$ is the only term in the above that depends on the model $f_q(x)$, the function that minimises the MSE also minimises the cross entropy.

If $p(y|x)$ is a normal distribution of the same form as Equation 6.6, and the standard deviations of both distributions are similar in size, with $\sigma_q = \sigma_p + \delta$ where $|\delta| \ll |\sigma_p|$, then we can expand the relative entropy (see Appendix B.1) to find that

$$D\big(p(y|x)||q(y|x)\big) \approx \frac{1}{2\sigma_p^2}(\mathrm{MSE}_{q,p} - \mathrm{MMSE}_p), \tag{6.8}$$

where MMSE_p is the minimum MSE that is achievable from data sampled from $p(x,y)$. Thus, for normally distributed errors, the relative entropy is proportional to the positive difference between the error achieved and the minimum possible error. This type of regret ensures that a misapplied model, in this case $f_q(x)$, achieves a higher error on average than the true model, $f_p(x)$. A more general expression relating relative entropy and regret for normally distributed errors has been recently discovered by Verdú[258].

We note relationships similar to the above, between mean absolute error and cross entropy can be found when the random errors have a Laplace distribution about the true model.

6.4 The transferability of structure-based models

We have shown that the PDF of structures and affinities for a particular protein-ligand complex, $p(x, y|\theta)$, is different for each molecular pair. Yet, structure-based scoring functions are fitted and applied to many different protein-ligand complexes. By using regression to train a model, one implicitly assumes that the complexes in a data set set have been sampled in a probabilistic manner. We denote the probability for selecting a complex θ for a particular data set as $\alpha(\theta)$. A scoring function is not trained on the complexes themselves, but on their structures and affinities sampled from the weighted sum

$$p_\alpha(x, y) = \sum_\theta p(x, y|\theta)\alpha(\theta), \tag{6.9}$$

where the sum is over all protein-ligand complexes and we have made the dependency of $p_\alpha(x, y)$ on $\alpha(\theta)$ explicit with a subscript. The optimal scoring function for complexes sampled from $\alpha(\theta)$ is encoded in the conditional PDF

$$\begin{aligned} p_\alpha(y|x) &= \frac{p_\alpha(x, y)}{p_\alpha(x)} \\ &= \frac{\sum_\theta p(y, x|\theta)\alpha(\theta)}{\sum_\theta p(x|\theta)\alpha(\theta)}. \end{aligned} \tag{6.10}$$

By inserting the PDFs defined in Equations 6.1 into the above, it is apparent that $p_\alpha(y|x)$ contains contributions from all the individual Boltzmann distributions of each complex, so that different protein-ligand sampling probabilities will, in general, result in distinct optimal models.

Using the above definitions, we now investigate the error of a scoring function that occurs when it is applied to a set of complexes sampled from $\alpha(\theta)$ but is trained on complexes sampled from a *different* distribution, denoted as $\beta(\theta)$. Following from Section 6.3, the bias incurred by applying the scoring function that is optimal on data sampled from $\beta(\theta)$, which is encoded in $p_\beta(y|x)$, to complexes sampled from $\alpha(\theta)$, is given by $D\big(p_\alpha(y|x)||p_\beta(y|x)\big)$. Our first main result (proven in the Appendix B.2) is that

$$D\big(p_\alpha(y|x)||p_\beta(y|x)\big) \leq D\big(\alpha(\theta)||\beta(\theta)\big). \tag{6.11}$$

Thus, the relative entropy between the complex selection probabilities is an upper bound to the bias of a misapplied scoring function. Scoring function bias is minimised when $\alpha(\theta) = \beta(\theta)$, so that the protein-ligand complexes in the training and test sets of a scoring function have been sampled from the same probability mass function. However, if $\alpha(\theta)$ and $\beta(\theta)$ do not overlap, then $D\big(\alpha(\theta)||\beta(\theta)\big)$ is unbounded, implying that scoring function error can be *arbitrarily* large. This can occur when the probability of finding any complex from the training set in the test set is zero, such as when a protein-specific QSAR model is applied to another protein.

A universally applicable structure-based model should, by definition, have the lowest possible error when applied to the widest conceivable range of protein-ligand complexes. If a training set is composed of a diverse range of proteins and ligands, we know from Equation 6.11 that for a scoring function's bias to be zero, the test set should be similarly diverse. In a typical virtual screen, however, scoring functions that have been trained on a diverse range of protein-ligand complexes are applied only to ligands binding to a *single* protein, implying different complex selection probabilities between the training and test sets and a potentially large bias.

6.4.1 The errors of generalised structure-based models

To investigate the error of a universal scoring function, we consider M different proteins that one can use to construct data sets of protein-ligand structures x and affinities y. Protein-specific complexes are sampled according to $\alpha_i(\theta)$, $i = 1, 2, 3, ..., M$, with corresponding joint PDFs $p_i(x, y)$. Many samples from $p_i(x, y)$ results in a data set of bound structures and affinities of different ligands bound to protein i. To model the creation of a data set composed of a diverse range of protein-ligand complexes, we select which $p_i(x, y)$ to sample from with probability ω_i. Similar to Equation 6.9, the appropriate joint PDF for this diverse scoring function is given by the weighted sum

$$\begin{aligned} p_\omega(x, y) &= \sum_i \omega_i \, p_i(x, y) \\ &= \sum_i \omega_i \sum_\theta p(x, y|\theta)\alpha_i(\theta). \end{aligned} \tag{6.12}$$

The corresponding conditional PDF, $p_\omega(y|x)$, encodes for the optimal diverse scoring function. If M is sufficiently large and ω suitably broad, then $p_\omega(y|x)$ represents a 'universally' applicable structure-based model. However, it is important to note that this generalised optimality is defined only for its particular sampling probabilities.

From the previous section, we know that a scoring function that has been designed for a diverse range of protein-ligand complexes will have a non-zero regret when applied to a protein-specific data set. A relevant question is therefore, how large is the average bias incurred by applying a scoring function defined from $p_\omega(y|x)$ to complexes sampled from $\alpha_i(\theta)$? Our second main result (see the Appendix B.3), is that for an arbitrary scoring function encoded by $q(y|x)$,

$$\sum_{i}^{M} \omega_i \, D\big(p_i(y|x)||q(y|x)\big) \geq \sum_{i}^{M} \omega_i \, D\big(p_i(y|x)||p_\omega(y|x)\big), \tag{6.13}$$

meaning the generalised model has the lowest average bias over all the protein-ligand complexes sampled by ω. Although the universal model $p_\omega(y|x)$ is optimal over the all M proteins, we now show, perhaps counter-intuitively, that it is not optimal for each specific protein. In our information theoretic perspective (see Equation 6.4), the minimum error of the general model $p_\omega(y|x)$ over the M proteins is given by its conditional Shannon entropy, denoted $h_\omega(Y|X)$. Also, the minimum achievable error of an optimal protein-specific model $p_i(y|x)$ on protein i is denoted $h_i(Y|X)$. Our third main result (see the Appendix B.4 for the proof) is that

$$\sum_{i}^{M} \omega_i \, h_i(Y|X) < h_\omega(Y|X). \tag{6.14}$$

Hence, the minimum error of a generalised structure-based model is greater than the average minimum error for protein-specific models. Thus, on average, a scoring function targeted for a specific protein will outperform a scoring that has been designed for a diverse range of protein-ligand complexes. This result explains previously reported studies[252]. Equation 6.14 also shows that a universal model has a broader error distribution than the average single-protein model.

For a given model, there is an unavoidable trade-off between accuracy over a broad spectrum of complexes and accuracy for certain individual cases. It is important to highlight that the generalised PDF, $p_\omega(y|x)$ and the PDF for a specific data set $p_i(y|x)$ are, in general, different. As these conditional distributions encode their own optimal

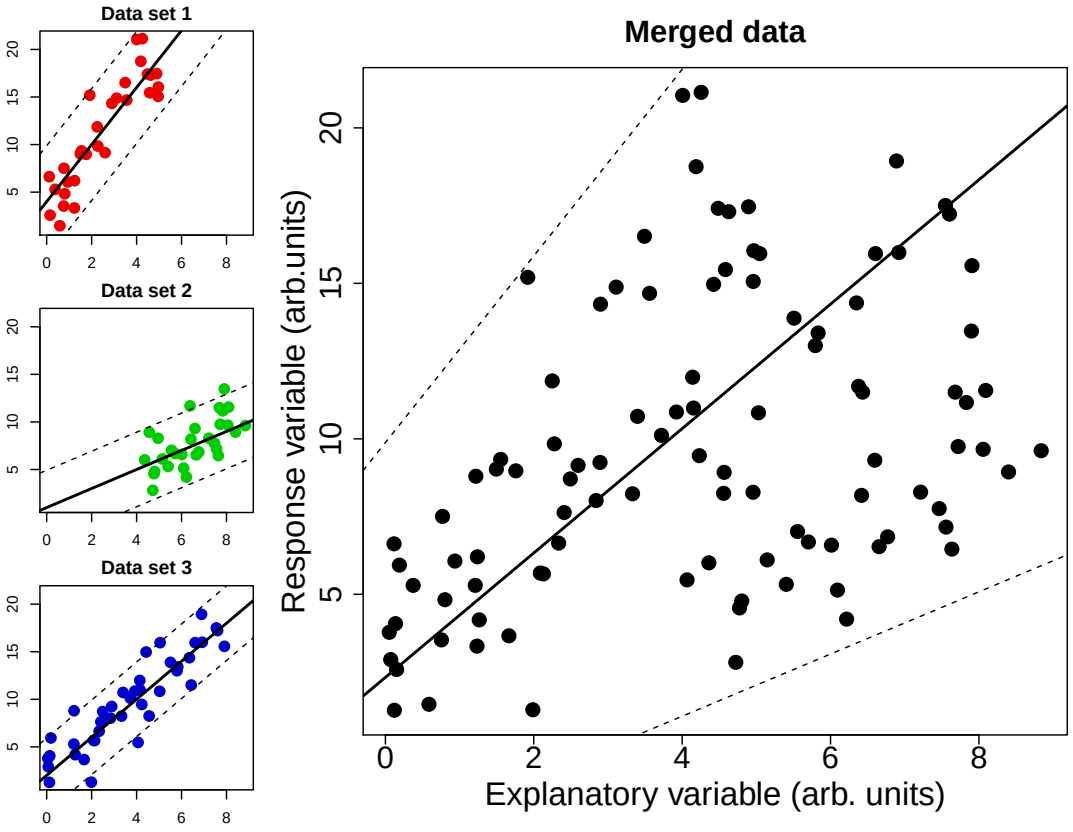


FIGURE 6.2: Synthetic data summarising the compromise between the generality and accuracy of a structure-based scoring function. The left column shows a fictional response variable plotted against a fictional explanatory variable for three different data sets. The data sets represent ligand binding data from three different proteins. The main plot shows the merged data from the three different sets. Each data set is governed by a different PDF, which is indicated by the true lines of best fit (black lines) and 95% confidence intervals (dashed lines). Like a universally applicable scoring function, the true line of best fit on the merged data has the lowest error on all the three data sets, yet its 95% confidence region is larger than the confidence regions on the individual data sets. Thus on average, the error of this general model is larger than the error for each specific data set. This is explained by Equation 6.14.

functional relationships between structure and affinity, the model that best predicts the binding affinities for many proteins may be very different from the best model for specific proteins.

6.4.2 The optimisation of scoring functions

The previous section showed that regret is integral to a generalised empirical scoring function. There is a similar trade-off in accuracy when optimising a scoring function for any set of complexes.

As described, the relative entropy quantifies the cost of misapplying a scoring function to a data set. This cost is a bias, and as it ensures that the minimum error is not achieved, it contributes to the regret of a model. The bias can be reduced for a particular sampling regime of protein-ligand complexes by re-fitting a scoring function for those complexes. However, when the regret for those complexes decreases, the regret between the scoring function and another sampling regime may increase. This occurs by virtue of the fact that different complex sampling probabilities results in distinct structure-affinity PDFs, and that the relative entropy is a convex function of two distributions. This means that a structure-based model that achieves the lowest possible error on a group of protein-ligand complexes will necessarily perform poorly on another group.

In conceptualising the practical implications of this, it is fruitful to consider two protein-specific conditional PDFs, $p_\alpha(y|x)$ and $p_\beta(y|x)$, as normal distributions with variances approximately equal to σ^2 and corresponding optimal scoring functions $f_\alpha(x)$ and $f_\beta(x)$ respectively. Following from Equation 6.8 and Heskes[259], we show in the Appendix B.1 that

$$D\big(p_\alpha(y|x)||p_\beta(y|x)\big) \approx \frac{1}{2\sigma^2} \int p_\alpha(x) \left(f_\alpha(x) - f_\beta(x)\right)^2 \, \mathrm{d}x. \tag{6.15}$$

This Equation implies that we can represent the theoretically optimal scoring functions $f_\alpha(x)$ and $f_\beta(x)$ as being embedded in a space that preserves the mean squared distances between them. Optimising a scoring function to a particular protein can then be considered as moving through this space towards the optimal model, increasing the distance from another model which is optimal on a different protein. A schematic diagram of this hypothetical scoring function space is shown in Figure 6.3.

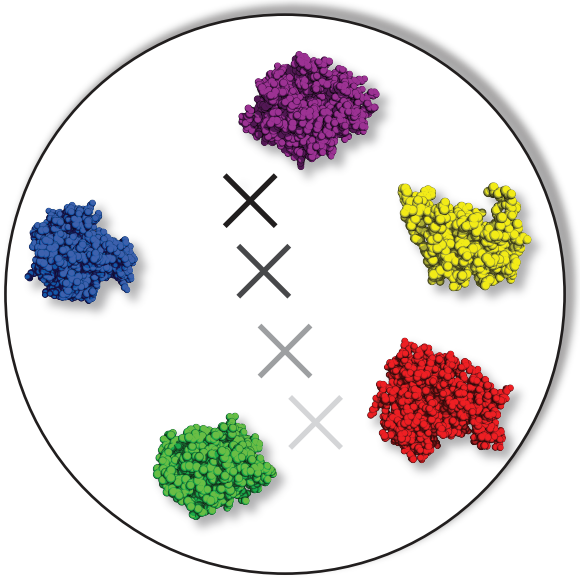


FIGURE 6.3: Schematic representation of a hypothetical scoring function space (circle). Every location in the space corresponds to a scoring function and each coloured protein represents the optimal model for a particular set of protein-ligand complexes. Beginning with a fitted model, shown as the lightest grey cross, that performs close to optimally on the green data set, its large distance from the purple model means that the bias on this set is also large. Optimising this model for the latter, shown by darkening gray crosses, increases the distance, and hence the error, from the green data set. Having a low error on one set of complexes necessarily means a scoring function performs worse on others. Using this graphical perspective, the scoring function that has the lowest bias on all five data sets can be represented as the position with the lowest average distance from all their optimal scoring functions, and would lie near the centre of the circle. As by definition this "general" scoring function is a finite distance away from each of the system specific scoring functions, it is not optimal on any of them.

Given that the above discussion and the upper bound in Equation 6.11, it is clear that transferring a scoring function that has been developed on a data set to new, previously unseen data can lead to large errors. This raises the following question: "is it possible to predict the error of a scoring function on new data?". Using a Bayesian formalism, Wolpert[260,261] proved that without knowing *a priori* which conditional PDF the test data obeys, all models had the same average error, so that when comparing two models, there are just as many test sets for which one model is 'superior' to the other as there test sets for which the same model is 'inferior'. Our analysis shows that this so called "no free lunch" theorem applies to structure-based models. While this result has been arrived at via formal analysis, it is noteworthy that the same conclusion has also been arrived at by Warren et. al.[25] after comparing different scoring functions in virtual

screening:

> "While we have demonstrated that virtual screening is successful, we have also shown that in the absence of prior knowledge about the protein target program performance was inconsistent across the target types evaluated. This inconsistency means that when there is an absence of knowledge about the target, one cannot predict a priori whether a particular program will be successful against the given target."

6.5 Scoring with missing information

The previous sections detailed the fundamental errors incurred when transferring an empirical scoring function from one data set to another and the errors incurred by generalised model. Another distinct source of error for structure-based scoring functions arises from the speed required for virtual screens. Soring functions typically utilise only a single snapshot of a bound protein-ligand complex to predict affinity, whereas in actuality, the binding affinity of a pair of molecules depends on both the bound and unbound ensembles of the protein-ligand pair. Clearly, the predictions are made in the absence of a great deal of information. In this section, we discuss how this lack of information impacts scoring function accuracy.

6.5.1 Regret and discarded data

Here, we analyse the regret incurred by discarding relevant information. As before, Y denotes affinity and X the structural snapshot used in scoring. We denote missing information as Z, which details the unbound ensemble as well as the bound ensemble not represented by X. As Y depends on the bound and unbound ensembles, the "true" conditional PDF, which encodes the most accurate scoring function, is $p(y|x,z)$. An arbitrary scoring function that only uses X to predict Y we denote as being encoded in $q(y|x)$, so

that the regret of using $q(y|x)$ instead of $p(y|x,z)$ is given by $D\big(p(y|x,z)||q(y|z)\big)$. We prove in Appendix B.5 that

$$D\big(p(y|x,z)||q(y|x)\big) \geq I(Y,X;Z) - I(X;Z), \tag{6.16}$$

where $I(Y,X;Z)$ is the mutual information between the data we don't have (Z) and the available structures (X) with affinities (Y), and $I(X;Z)$ is the mutual information between the available structures (without affinities) and the missing data. Mutual information quantifies the reduction in uncertainty in one random variable that is achieved by knowing the other. It can be zero only when the variables are independent and otherwise is positive. For a data set of rigid body, 'lock and key' binders, $I(X;Z)$ will be very high, as a single structure of the complex is representative of both the bound and unbound conformation ensembles. In this case, Equation 6.16 implies that the average regret of a structure-based scoring function applied to such a system can be low. Conversely, if the dynamics of the bound state differ substantially from those of the unbound state, and a single structure is not archetypal of the bound conformations, we would expect a correspondingly large error. In general, Equation 6.16 can be interpreted as quantifying the penalty for modelling with missing information. Contrary to the famous phrase, what we don't know *can* hurt us, at least where predictive accuracy is concerned.

Equation 6.16 also implies that for a given set of known structures of a single protein bound to many different ligands, the relative effect the ligands have on the protein's activity can be predicted with a lower regret, on average, than can their absolute binding affinities. This is because the activity is a function of the bound dynamics of the ligand and protein, while affinity depends on both the bound and unbound ensembles of the protein and ligand. While the unbound information is missing, the free energy of the

unbound protein is a constant, and will not affect a rank ordering of the affinities or activities. The free energy of the unbound ligand, however, is a variable, and hence not knowing it will hinder the accuracy of the predicted binding affinity, though not the predicted activity.

6.5.2 Missing information in forcefield-based scoring functions

While empirical scoring functions have been the primary focus of our analysis, forcefield-based scoring functions provide a clear example where missing information hinders accuracy. Forcefield-based scoring functions approximate the binding free energy of a complex with an estimate of the interaction energy of the protein-ligand complex. In techniques such as the relaxed complexed scheme, the calculated interaction energy is averaged over multiple frames from molecular dynamics simulations of the bound protein-ligand complex[262–265]. Multiple crystal structures have also been used to calculate average interaction energy for affinity estimation[266,267]. The interaction energy is an approximation of the total change in potential energy when moving from the unbound ensemble to the bound ensemble, denoted $\Delta E_{u\to b}$. The change in potential energy averaged over the bound state of a protein-ligand complex is denoted as $\langle \Delta E_{u\to b} \rangle_b$. By expanding the relative entropy between the Boltzmann distributions of the bound and unbound states, it is evident that $\langle \Delta E_{u\to b} \rangle_b$ is in fact a biased estimator for the binding free energy of a complex. For simplicity, the change in Helmholtz free energy is considered, denoted $\Delta F_{u\to b}$, which is valid for a a system of constant volume, temperature and particle number. The relative entropy of the bound and unbound probability distributions is given by

$$k_B T \, D(p_B(\mathbf{r}^N|\lambda_b)||p_B(\mathbf{r}^N|\lambda_u)) = \Delta F_{u\to b} - \langle \Delta E_{u\to b} \rangle_b, \tag{6.17}$$

The above relative entropy quantifies the amount of information $p_B(\mathbf{r}^N|\lambda_b)$ does not contain about $p_B(\mathbf{r}^N|\lambda_u)$, and Equation 6.17 states that the bias of an average interaction energy estimator is determined by the degree of overlap between the bound and unbound probability distributions. A high degree of overlap can occur when two binding partners have very similar conformational flexibilities when they are bound to when they are unbound. Thus, in a similar manner to Equation 6.16, average interaction energy methods will have a low bias when applied to rigid body binding. Forcefield-based scoring functions further assume that $\langle \Delta E_{u\to b} \rangle_b$ is dominated by single state.

While complimentary to Equation 6.16, Equation 6.17 relates missing information and error for a single complex, and not over a data set. If $\langle \Delta E_{u\to b} \rangle_b$ is used to estimate $\Delta F_{u\to b}$ for multiple complexes, $D\big(p_B(\mathbf{r}^N|\lambda_b)||p_B(\mathbf{r}^N|\lambda_u)\big)$ is a source of variance for the model. This is because the KL divergence does not remain constant for each binding pair, as each bound complex could contain different amounts of information of the unbound ensemble. For congeneric ligands, this variance will be small if the dynamics of the protein and ligands all change similarly upon binding.

While we found Equation 6.17 independently, this relationship has been highlighted in a study concerning image analysis[268]. Equation 6.17 is also analogous to a equation in non-equilibrium fluctuation theory[269].

6.6 Verification of analytical results

Chapter 5 described the development and testing our own generalised scoring functions. The most robust of our models was linear with five explanatory variables which described only protein-ligand interactions. Here, we re-assess this scoring model in light of our analytical results regarding universal scoring functions.

We found that the accuracy of our scoring functions were heavily dependent on the data set they were applied to. This ultimately follows from the "no free lunch" theorem for supervised learning[260,261]; a scoring function that performs well on one sampling regime of protein-ligand complexes necessarily means it will perform poorly on others. Also, we know from Equation 6.11 that the regret of a misapplied scoring function can be arbitrarily large if the complexes in the training and test sets have been sampled in a significantly different manner. We further illustrate this difficulty in scoring function optimisation by plotting the mean absolute error of our simplest linear model for each free parameter pair used in the elastic net regularisation (see Chapter 5 Section 5.2.2). Changing these parameters can be thought of as moving through a hypothetical scoring function space as depicted in 6.3 towards the location of the optimal protein-specific model. The relative mean absolute error of the scoring function for each free parameter pair on each data set are shown in 6.4. No single choice of parameter pair yields the minimum error on all of the sets, and 6.4 shows that optimising the scoring function for a data set can increase the error on others.

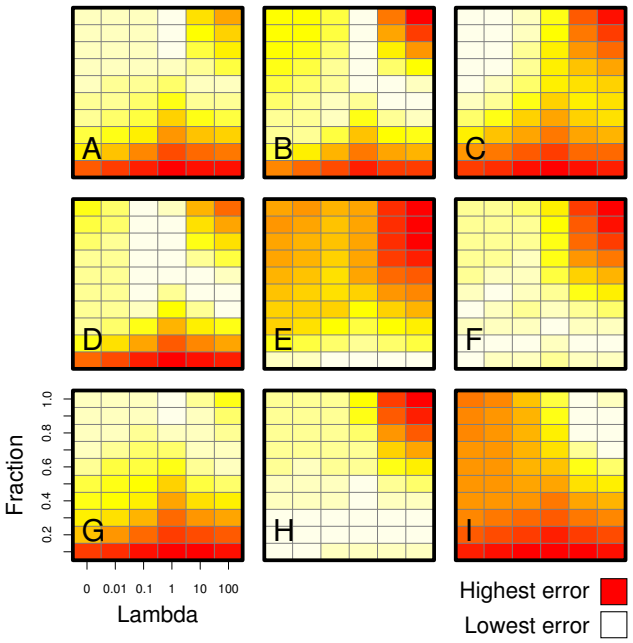


FIGURE 6.4: Grid of heatmaps showing the mean absolute error for our own scoring function when applied to data sets A-I (see Table 5.2). Data set A is composed of many different protein-ligand complexes while data sets B-I are single-protein data sets. Each heatmap shows the relative error of the model on the data set as the two free parameters of the scoring function – trained using elastic net regularisation – are varied (x and y-axis). The color gradation indicates parameter-pair choices that give rise to the lowest (white) through to the highest (red) error for that data set. Coloring by relative error highlights that no single parameter-pair choice achieves the lowest error on all nine data sets, so that a model that has the lowest average error over all data sets will not be the best on each individual data set. To compare the absolute values of the errors, Figure A.1 of the Appendix utilises an absolute coloring scale.

Equation 6.13 shows that, for a particular sampling regime, the conditional PDF that encodes the scoring function with the lowest error over a diverse range of complexes also has the lowest average bias than any other. Yet Equation 6.14 shows that this comparatively low bias is compensated by an intrinsic error that is larger than the average minimum error for specific protein models. This implies that protein-specific scoring functions will have a lower error when applied to their respective protein than the general model. Similarly, misapplying a protein-specific model to the wrong protein will have a higher average error than a generalised scoring function. To approximate protein-specific scoring functions, our most robust linear scoring function was fitted to the single protein data sets shown in Table 5.2 in Chapter 5. The elastic net fitting

procedure was used again, and the free parameters were optimised using leave-one-out cross validation to create 8 *single-protein* scoring functions. The average error from cross validation of each single-protein model is 1.2 kcal/mol compared to an average error of 1.8 kcal/mol of the diverse scoring function. By applying each single-protein scoring functions to the other data sets, we found the average misapplied error to be 2.7 kcal/mol. The relative sizes of each of these errors are in complete agreement with our analytical results. The mean absolute errors of these models on each data set are shown in 6.5.

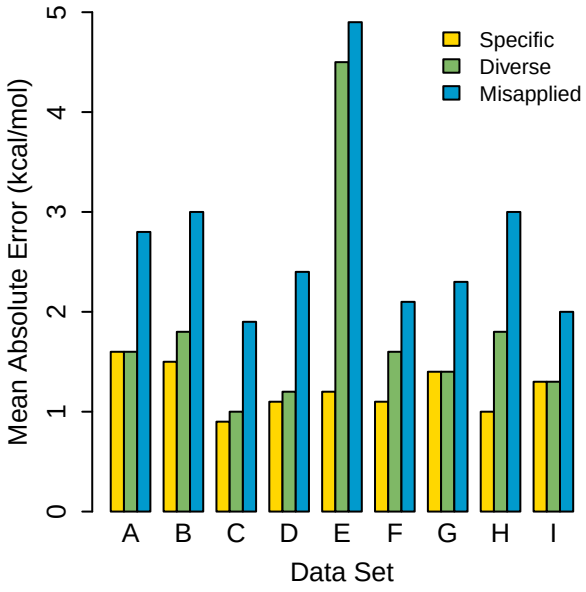


FIGURE 6.5: The mean absolute errors of our scoring model on our test sets (see Table 5.2) when fitted in three different ways. Yellow bars show the cross validation error of the model when it is fitted to each specific data set; green bars show the error of the model when it is fitted to a diverse range of protein-ligand complexes; blue bars show average errors of the protein-specific scoring functions when misapplied to that data set. In agreement with our mathematical analysis, the error of a protein-specific scoring function is on average less than the error of a general scoring function, which itself is more accurate than the average error of a misapplied protein-specific scoring function.

By assuming that the cross validation error of the single-protein models is the minimum error achievable for our scoring function descriptors, we can approximate the regret of the diverse scoring function on a single-protein data set by the difference between the error it achieves and the cross validation error. On the carbonic anhydrase II data set

(E), the diverse scoring function has an exceptional large regret of around 4 kcal/mol. Removing this data set from the analysis, the diverse-scoring function has an apparently encouraging average regret of 0.3 kcal/mol. In actuality, a truly protein-specific scoring function would be have to be designed from the bottom up, and may include descriptors designed especially for the protein, have a particular functional form and have the method of regression chosen after experimentation. The "protein-specific" scoring functions used above are only recalibrations of the diverse scoring function. Thus, the regret we report of our diverse model is likely to be larger when truly protein-specific models are considered.

We attribute the large error on carbonic anhydrase II to the relative infrequency of complexes with metal-ligand interactions appearing in the diverse training set. Carbonic anhydrase II, has a catalytic zinc ion in its binding site that interacts with its inhibitors. In the diverse training set, only 37 out of the 207 ligands are within a hydrogen bonding distance to a metal ion. Thus, the data set contains relatively little information about metallic interactions and a correspondingly large regret for these type of complexes is to be expected. Metalloproteins are notoriously difficult to account for in scoring functions, and further optimisation or extra protocols are typically required for specific systems[270–274]. The substantially lower error our scoring function attains when recalibrated for carbonic anhydrase II is indicative of this general trend. As our analytical results show, the accuracy of an empirical structure-based scoring function depends on the degree of informational overlap between the training and test sets, such that specific scoring functions have a lower error on average than generalised models.

6.7 Summary

By formally analysing the structure-based modeling process, we have conclusively proven that protein-specific scoring functions will on average achieve a lower error than the very best universal model. We have shown how training and applying a scoring function on different sets of protein-ligand complexes can result in an arbitrarily large error, and that a model which performs optimally on one set of complexes necessarily performs poorly on another set by virtue of the 'no free lunch theorem' for supervised learning.

Our results follow from the fact that data sets of protein-ligand structures and affinities are, in general, governed by distinct probability distributions, so that there may be a cost in transferring empirical scoring functions between data sets. This cost is a bias that contributes to the regret of a model. We employed an information theoretic analysis to demonstrate that this error is independent of the way protein-ligand interactions are modeled and is a property of the data itself. Thus, error via bias is fundamental to the nature of protein-ligand scoring. While previous research into the sources of scoring function error has focused on errors from energetic calculations[254], bias-derived error has remained unreported and explains the variability of scoring function performance on different protein data sets. Given that model regret is intrinsic to protein-ligand scoring, it remains of paramount importance to be able to estimate its magnitude in a predictive setting. Any further understanding regarding the limits of scoring function accuracy is vital for a more informed and efficient synergy between theoretical affinity predictions and experimentally driven drug development.

Chapter 7

Conclusions

7.1 The prediction of protein-ligand binding affinities

Chapter 6 formally proved the existence of errors which occur when transferring empirical models to and from proteins and ligands, and that generalisation comes at the expense of accuracy. The outlined theory does not, however, tell us the numerical value of a performance measure, such as MSE, that the very best scoring function could achieve. Have current models already hit the fundamental limits already? If not, how much more can we expect accuracy to improve? When addressing these questions, we must be careful to distinguish between models that are designed to be of a universal nature, and models that are for specific systems.

Ultimately, the best achievable correlation between predicted and measured affinities is set by experimental error. In recent years, there has been a growing appreciation for discrepancies between different methods and researchers, and the impact these can have on empirical models[137,275–278]. By analysing over two-thousand samples of affinity measurements that were repeated by different research groups, Kramer et. al. reported

the mean absolute difference between measurements to be 0.6 kcal/mol[278]. They calculated that a *perfect* scoring function, which predicts "true" affinities, tested on a diverse protein-ligand data sets, such as CSAR or PDBbind, would have an $R^2 \approx 0.8$ due to experimental error. If a scoring function is as accurate as experiment, the correlation drops, with correlation $R^2 \approx 0.7$. For comparison, our gradient boosted tree models in Chapter 5 all achieved $R^2 = 0.65$ on our CSAR diverse validation set. Spearman rank correlation coefficients were higher, with all models achieving correlations just under $\rho = 0.80$. Similar accuracies have also been reported by other scoring functions on the diverse PDBbind test set[235]. These findings suggest that universal scoring functions are already close to their maximum performance.

The focus on universal scoring functions, however, may present somewhat of a distraction in computer-aided drug design. In virtual screening, for instance, scoring functions are used to predict the affinities of compounds against a *single* protein. Chapter 6 demonstrated that with empirical scoring functions, protein-specific models can achieve a lower error on average than universal ones, indicating that efforts should concentrate on developing robust specific models from whatever data is available. Problems with universal scoring functions are not only limited to empirical models. Knowledge-based scoring functions, for instance, are based on atomic distance frequencies dervived from diverse ranges of protein-ligand complexes, so are likely to suffer from the same inadequacies as universal empirical models, although a formal analysis is required to verify this. Chapter 6 additionally showed that forcefield-based scoring functions are limited for use with rigid-body binders and congeneric ligands.

Benchmark test sets for scoring functions have traditionally been composed of a wide range of protein-ligand complexes[207], reflecting the field's emphasis on universal models. Yet, the poor performance of universal models on specific proteins, has fuelled a growing movement to concentrate more on protein-specific test sets[28,31,137,245]. The

theory and approaches underlying scoring functions and QSAR models has remained largely unchanged since the mid 1990s[7], with non-linear empirical scoring functions first appearing in 1996[244]. One can hope that this evolving mindset will encourage novel approaches to scoring.

New scoring methods may draw inspiration from the field of *ab initio* protein folding, in which the three-dimensional structures of proteins are predicted from sequence alone. As it is often assumed that folded proteins are in a state of minimum free energy, the problem shares an interesting duality with structure-based affinity prediction: scoring functions attempt to predict free energies from estimated structures, while folding models try to predict structures from estimated free energies. Evolutionary sequence covariance analysis[279], made possible by the large number of protein sequences available, provides possible contact information between protein residues. The covariance analysis, while prone to errors, has been coupled with molecular mechanics, fragment-based approaches and statistical potentials to drive recent improvements in accuracies, particularly with membrane protein structure prediction[280,281]. It is the successful combination of *independent* sources of information that protein-ligand scoring has the most to learn from. As Equation 6.16 shows, the inclusion of extra information reduces the lower bound to a model's error. Currently, binding free energy estimates are compartmentalised: ligand-similarity methods, structure-based scores, and rigorous molecular simulation methods have not yet been combined into a single protocol. It is likely that combining all such knowledge will result in improved affinity estimates. Bayesian methods provide an elegant framework in which to combine different sources of information, and therefore, are likely to be of even greater benefit for affinity prediction than they currently are[282,283]. For example, the predictions from a scoring function could be used to form a prior distribution for a rigorous alchemical method, such as exponential averaging. The prior

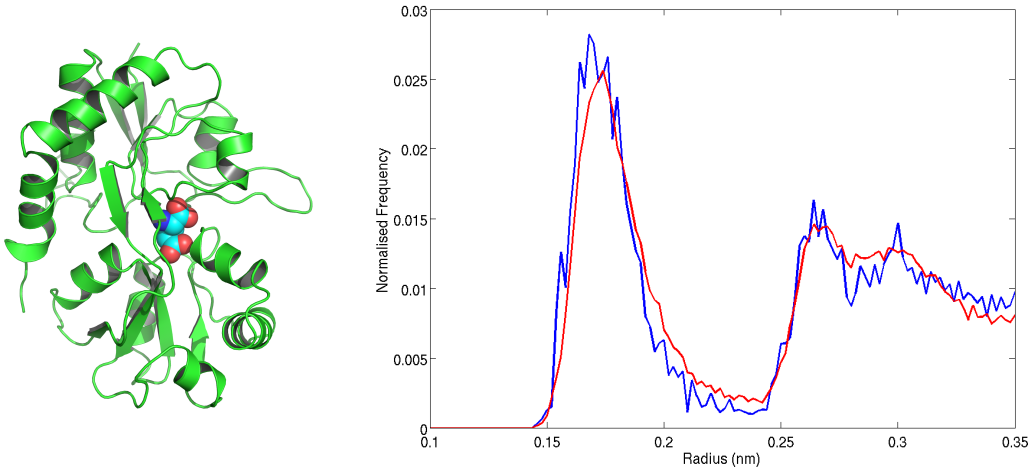
distribution would act to stabilise the rigorous affinity estimate when only a small number of samples are used, potentially allowing for shorter simulations. To the best of our knowledge, this has yet to be investigated.

Despite the limitations in universal scoring functions, creating a protein-specific model is very difficult when there is little experimental data available on the target. The development of protein class or family specific models may be a sufficient compromise between generality and accuracy. When dealing with small data sets, these generalised models can also provide the prior distributions on the regression coefficients for Bayesian regression, as it is well established that Bayesian regression greatly stabilises fitted models in such cases[144].

7.2 Scoring water in protein-ligand interactions

This thesis began by hypothesising that a more detailed understanding of water in protein-ligand complexes can be pursued as a route to improve rapid affinity estimation. How valid is this hypothesis in light of the preceding chapters? What are the avenues for future work?

Firstly, it is worth noting that an empirical scoring function which calculates water binding energies, such as the one developed in Chapter 4, can expect to be more accurate than a universal scoring function for protein-ligand complexes. This is because a water scoring functions is an example of a specific model, being designed only for one ligand. However, experimentally measuring the binding energy for a water binding to a particular site is extremely difficult, if not impossible, meaning that water scoring functions can only be trained on binding energies calculated via rigorous simulation methods, which can be highly sensitive to the water model used[284].



(A) (B)

FIGURE 7.1: (A) The ligand binding domain of GluR2 is an example where water plays important roles in ligand binding [154,285]. The figure shows glutamate bound, PDB code 1FTJ. (B) Radial distribution functions (RDFs) calculated from molecular dynamics simulations of water around one of glutamate's carboxylate oxygen atoms. Simulation details are as in Appendix A.1, with glutamate in water simulated for 100 ns and the GluR2-glutamate complex for 10 ns. The RDF calculated from the binding site (blue) retains the same shape up to the second solvation shell as in bulk water (red). As a result, this atom is likely to pay less of a desolvation penalty when binding to the protein.

Effects due to water contribute significantly to the affinity of a protein-ligand complex. In the water-based scoring functions of Chapter 5, the carbon-carbon contact term was by far the most important for predicting affinity (see Figure 5.6), emphasising the importance of the hydrophobic effect. Preliminary investigations using molecular dynamics of glutamate in bulk water and bound to the ligand binding domain of GluR2 indicate that bridging water molecules act to reduce the desolvation cost of transferring ligand chemical groups from bulk water (see Figure 7.1). The importance of a ligand holo water hydrogen bond-like term suggests that the water-based scoring function partially captured this effect. As such, treating explicit water effects from the ligand's perspective, such as in the docking method by Forli [175], merits further investigation. Yet, exploiting novel explicit water effects for more accurate predictions was our primary aim, and in this regard, the water-based scoring function was unsuccessful. Partly, this is because the water-based scoring function was designed to be universal, and as Section 6.4.1

demonstrates, a universal model can be considered as an average over protein-specific models. As a result, any interesting protein-specific water interections may have been smeared out. Utilising the water models for particular test cases may be more fruitful.

While a more detailed model, such as accounting for binding site and ligand desolvation, may have a higher fidelity to true underlying phenomena, the larger number of parameters that may be required, and the uncertainty in the input values, conspire to increase the error above a much simpler model: the hallmark of the bias-variance trade-off (see Chapter 2, Section 2.3). The binding modes of ligands, the locations of water molecules and binding site side-chain rotomers are inherently uncertain, owing to their thermodynamic entropy (see Equation 3.1). These uncertainties propagate via Equation 5.10 to ensure that scoring functions inherently suffer from high variances. Regression techniques, such as elastic net regularisation used in Chapter 5, are specifically designed to lower variance at the expense of a higher bias. This explains the comparatively good performance of our simplest scoring function (see Tables 5.3 and 5.5), and why recent testing of the AutoDock 4 scoring function saw the greatest improvement with a biased regression method, rather than with advanced electrostatic charge models[230]. Therefore, robust regression techniques – which can often be formulated as types of Bayesian regression – should always be used when training scoring functions.

7.3 Strategies and future directions for virtual screening

Chapter 6 showed that there does not exist a "true" or universally "best" model to predict binding affinities at the empirical level, which in itself is sufficient to help inform a new strategy for high-throughput virtual screening. The typical strategy in virtual screening is a linear one: a scoring function is chosen, a virtual screen is performed, the top predictions are experimentally tested, and the verified binders are taken forward

for hit-to-lead development. Acknowledgement that the starting scoring function is not optimal suggests that an iterative approach may yield more hits. Instead of one round of virtual screening, multiple virtual screening rounds could be performed. After each round, the top hits are experimentally tested and the scoring function is modified to account for the new data. A virtual screen is performed again, and the new predictions experimentally tested. The process could be repeated as many times as resources allow. The scoring function – rather than being a fixed model – is constantly updated to account for the new data from the experimental assays. In terms of the fitting methodology, Bayesian regression techniques may be particularly useful. In this case, the regression coefficients of a universal scoring function could form the centres of prior distributions. As new experimental data is obtained, the distributions are updated via Bayes' theorem. At no point in the iterative process can the scoring function be declared optimal. Instead, it can be said to capture the current state of knowledge about the target protein and its interactions with the tested compounds. This procedure, while more costly than the current linear process, is automatable, and could be made cheaper by virtually screening and experimentally testing fewer compounds on each iteration than in the current linear process. Nevertheless, the additional effort can more easily be envisaged for a pharmaceutical company than in a small biotechnology firm or academic group.

When only a single virtual screen is feasible, as discussed, the use of a protein-specific scoring function is preferable. However, as there is always a cost to generalisation, models developed for specific congeneric series will be more accurate still. Clearly, this requires sufficient experimental data to train the models. When there is no experimental binding data available, one could construct a scoring function from other protiens that are homologous, or have structurally similar binding sites that *do* have experimental binding data. For instance, if the natural substrate of the protein is adenosine triphosphate, then a scoring function could fitted to data from other well studied proteins that

also bind to adenosine triphosphate. While the predicted hits from such a screen may not be specific for the target, the generated hits may be adequate for hit-to-lead optimisation. If there is no homology in sequence or structure with other proteins, a universal model must be used, as it best reflects the lack of target-specific knowledge.

The results presented in Chapter 6 also suggest that the manner in which compounds from a virtual screen are selected for experimental testing can be improved. Equation 6.14 and Figure 6.2 show that generalised scoring functions have a wide error spread because they encompass many sub-groups, each with their own optimal model, and with the minimum error within a subgroup being no larger than that of the generalised model. Thus, following a virtual screen, rather than testing the top hits from the whole compound data set, one should test the top hits from within each sub-group of compounds, as the relative affinities of a sub-group would have been predicted with greater confidense. Compound sub-groups, which may not be known *a priori*, can be estimated by clustering the ligands based on chemical similarity. This strategy provides a potential way to combine ligand-based and structure-based methods, and may have the added benefit of experientially testing a wider array of chemical space, which would better inform future models.

Throughout this thesis, different software and models have been developed to aid computer-aided drug design and virtual screening. WaterDock was developed to predict the locations of water molecules in protein binding sites, and classifiers were also developed to predict the role of water. Both of these tools have been freely disseminated to the research community. The simplest of the scoring functions that were developed in Chapter 5 may also be of use to the community, particularly because it has been fitted using a low-variance technique, and as dicussed, scoring functions suffer from high variances. Our intention is to release this following the publication of the work in Chapter 5. While

the speed required of scoring functions means that their accuracy will inevitably be compromised, it may be most fruitful to develop better strategies, rather than to attempt to create ever more accurate models, for virtual screens. In particular, the iterative virtual screening strategy and clustered hit selection described directly above merit further research.

7.4 Closing remarks

In this final chapter, the amalgamation of different affinity prediction methods, incorporating all prior knowledge, has been mooted as a route to improve accuracy. Yet, for the reasons discussed above, the addition of explicit water detail into the traditional empirical scoring function framework generally suffered from higher errors than simpler models. Thus, it is the *way* in which new information is incorporated into affinity models that is of central importance. The framework developed in Chapter 6 could be used to reappraise affinity prediction from first principals, so that in time, the theory underpinning future scoring functions can be of equal rigour to the exact methods based on statistical mechanics. However, unlike rigorous free energy calculations, fast methods do not have the luxury of sampling many molecular conformations, which would reduce the variance of estimates. Uncertainty, therefore, is an integral part of scoring, and I firmly believe that charting this unpredictable landscape will enable us to elevate rapid affinity calculations for the betterment of drug discovery.

Appendix A

Details of protocols and numerical results

A.1 Molecular dynamics simulation details

The unliganded structures of heat shock protein 90, PIM1 kinase and penicillopsin were selected from Table 3.1 to further validate water docking with Vina against molecular dynamics (MD) simulations. These structures were prepared and simulated using the GROMACS version 4.5.3[286] with the OPLSAA forcefield[215]. Bond lengths and angles were constrained using the LINCS algorithm[287] and the forces were integrated using a timestep of 0.002 ps. The nearest neighbor list was updated every 10 steps. Simulations were carried out in a cubic box with periodic boundary conditions, whose walls were initially 10 Å away from the proteins. The boxes were solvated with the TIP4P[288] water model and the systems were neutralized by replacing water molecules with sodium and chlorine ions up to a concentration of 0.15 M. The temperature was kept constant at 300 K using the Noose-Hoover thermostat[289,290] with a time constant of 0.1 ps. Pressure

was maintained at 1 atm using the Parrinello-Rahman barostat[291] with isotropic coupling and a time constant of 1 ps and compressibility of 4.5 × 10-5 bar^{-1}. Long range electrostatics were treated using the particle mesh Ewald[292] method with a cutoff of 10 Å.

The energy of the systems were minimized using the steepest decent algorithm and was then subject to a 200 ps simulation where the heavy atoms of the protein were restrained with a spherical harmonic potential with force constant of 1000 kJ/mol/nm^2. Finally, 10 ns unrestrained simulations were performed.

A.2 WaterDock OppA test set

PDB code	Resolution	Ligand
1JET	1.2	KAK
1JEU	1.3	KEK
1JEV	1.3	KWK
1B4Z	1.8	KDK
1B5I	1.9	KNK
1B32	1.8	KMK
1B3F	1.8	KHK
1B46	1.8	KPK
1B51	1.8	KSK
1B58	1.8	KYK
1B5J	1.8	KQK
1B9J	1.8	KLK
1QKA	1.8	KRK
1QKB	1.8	KVK

TABLE A.1: The X-ray crystal structures of OppA used to test set the water placement method against AcquaAlta[169].

A.3 Water docking protocols

Method				Results		
N^o of docking runs	Vina Exhaustiveness	Clust. method	Clust. cutoff (Å)	TP (%)	FP (%)	Mean error (Å)
3	20	Complete	1	90	35	0.74
3	20	Complete	1.6	90	28	0.78
3	20	Complete	1.9	86	25	0.79
3	20	Ward	1	91	35	0.74
3	20	Ward	3	89	27	0.77
3	20	Ward	6	80	19	0.79
1	10	S.L.*	1.6	77	25	0.81
1	20	S.L.	1.6	78	22	0.76
3	10	S.L.	1.6	85	27	0.74
3	20	S.L.	1.5	83	26	0.78
3	20	S.L.	1.6	86	23	0.78
3	20	S.L.	1.7	83	23	0.78
3	20	2 × S.L.	0.5 and 1.6	87	19	0.78

Table A.2: Results of different water docking methods on the Table 3.2 structures. *Single-linkage. The true positive rate (TP) is equal to the number of consensus water molecules predicted, and FP denotes false positive rate. The last method listed is the one chosen used for the final WaterDock protocol.

A.4 Scoring function results

Data set	MAE*	ρ	R^2
A	1.6	0.76	0.57
B	1.8	0.29	0.08
C	1.1	0.71	0.45
D	1.2	0.50	0.20
E	4.8	0.37	0.41
F	1.7	0.62	0.32
G	1.5	0.76	0.57
H	1.7	0.17	0.20
I	1.4	0.70	0.60

Table A.3: Table showing three performance measures of the simplest linear model (explanatory variable Set 1) on the validation and test sets (see Table 5.2). *Mean absolute error in kcal/mol. To be used with reference to Figure 5.2.

Data set	MAE*	ρ	R^2
A	1.5	0.80	0.65
B	1.8	0.25	0.14
C	0.9	0.78	0.59
D	1.3	0.18	0.14
E	4.9	0.34	0.28
F	1.8	0.25	0.10
G	1.6	0.74	0.55
H	1.3	0.14	0.13
I	1.7	0.39	0.27

TABLE A.4: Table showing three performance measures of the gradient boosted tree model with explanatory variable Set 1 on the validation and test sets (see Table 5.2). *Mean absolute error in kcal/mol. To be used with reference to Figure 5.2.

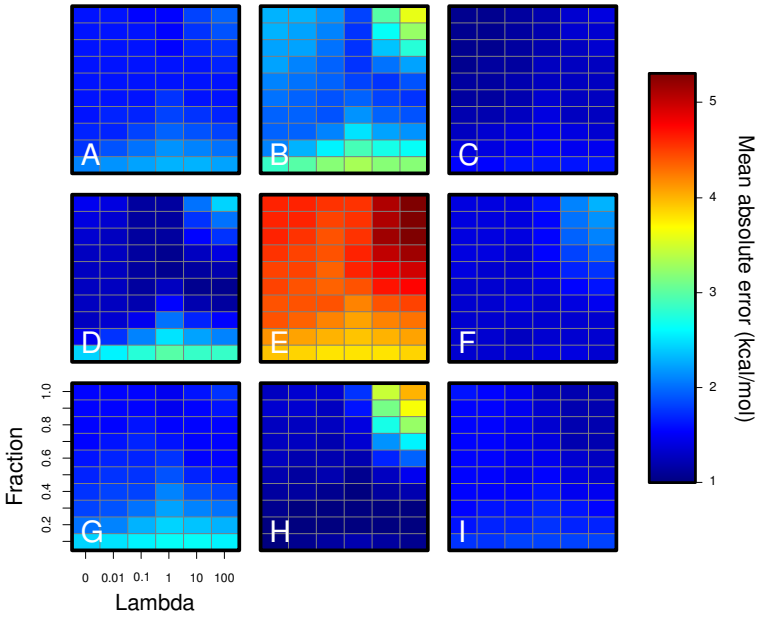


FIGURE A.1: Grid of heatmaps showing the mean absolute error for our own scoring function when applied to data sets A-I (see Table 5.2). Each heatmap shows the absolute error of the model on the data set as the two free parameters of the scoring functions are varied (x and y-axis). This figure supplements Figure 6.4 of the main text, which shows the relative errors on each data set.

Appendix B

Proofs of main analytical results

B.1 Proof of Equations 6.8 and 6.15

If one applies a scoring function that is optimal for a data set of structures and affinities sampled from $p_\beta(x, y)$ to a data set sampled from a different distribution $p_\beta(x, y)$, then the bias incurred by this model is given by

$$D\big(p_\alpha(y|x)||p_\beta(y|x)\big) = \iint p_\alpha(x, y) \ln \frac{p_\alpha(y|x)}{p_\beta(y|x)} \, \mathrm{d}x \, \mathrm{d}y. \tag{B.1}$$

From Equation 6.6, it follows that

$$p_\alpha(y|x) = \sqrt{\frac{1}{2\pi\sigma_\alpha^2}} \exp\left(\frac{-(y - f_\alpha(x))^2}{2\sigma_\alpha^2}\right), \tag{B.2}$$

where $p_\beta(y|x)$ is similarly defined. We investigate the case when the variances are similar in magnitude, and set $\sigma_\beta = \sigma_\alpha + \epsilon$ where $|\epsilon| \ll |\sigma_\alpha|$, so that the relative entropy

$$
\begin{aligned}
&D\big(p_\alpha(y|x)||p_\alpha(y|x)\big) = \frac{1}{2}\ln\left(\frac{\sigma_\alpha^2 + 2\epsilon\sigma_\alpha + \epsilon^2}{\sigma_\alpha^2}\right) \\
&+ \frac{1}{2(\sigma_\alpha^2 + 2\epsilon\sigma_\alpha + \epsilon^2)} \iint p_\alpha(x,y)(y - f_\beta(x))^2 \,\mathrm{d}x\,\mathrm{d}y - \frac{1}{2\sigma_\alpha^2}\iint p_\alpha(x,y)(y - f_\alpha(x))^2 \,\mathrm{d}x\,\mathrm{d}y \\
&\approx \frac{1}{2\sigma_\alpha^2}\iint p_\alpha(x,y)\,(y - f_\beta(x))^2 \,\mathrm{d}x\,\mathrm{d}y - \frac{1}{2\sigma_\alpha^2}\iint p_\alpha(x,y)\,(y - f_\alpha(x))^2 \,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{2\sigma_\alpha^2}(\mathrm{MSE}_{\beta,\alpha} - \mathrm{MMSE}_\alpha), \qquad \text{(B.3)}
\end{aligned}
$$

where the second line follows from the relative size of ϵ to σ_α and MMSE_α is the minimum mean squared error achievable on $p_\alpha(x,y)$ and $\mathrm{MSE}_{\beta,\alpha}$ is the mean squared error of $f_\beta(x)$ applied to $p_\alpha(x,y)$. As well as the regret, the relative entropy is also proportional to the mean squared distance between the true and estimated scoring functions. To show this, we start from Equation 6.8 so that we have

$$
\begin{aligned}
&D\big(p_\alpha(y|x)||p_\alpha(y|x)\big) = \frac{1}{2\sigma_\alpha^2}(\mathrm{MSE}_{\beta,\alpha} - \mathrm{MMSE}_\alpha) \\
&= \frac{1}{2\sigma_\alpha^2}\iint p_\alpha(x,y)\Big[(y - f_\beta(x))^2 - (y - f_\alpha(x))^2\Big] \,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{2\sigma_\alpha^2}\iint p_\alpha(x,y)\Big[y^2 - 2yf_\beta(x) + f_\beta(x)^2 - y^2 + 2yf_\alpha(x) - f_\alpha(x)^2\Big] \,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{2\sigma_\alpha^2}\int p_\alpha(x)\int p(y|x)\Big[-2yf_\beta(x) + f_\beta(x)^2 + 2yf_\alpha(x) - f_\alpha(x)^2\Big] \,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{2\sigma_\alpha^2}\int p_\alpha(x)\Big[-2f_\beta(x)\int p(y|x)y \,\mathrm{d}y + f_\beta(x)^2 + 2f_\alpha(x)\int p(y|x)y \,\mathrm{d}y - f_\alpha(x)^2\Big] \,\mathrm{d}x \\
&= \frac{1}{2\sigma_\alpha^2}\int p_\alpha(x)\Big[-2f_\beta(x)f_\alpha(x) + f_\beta(x)^2 + 2f_\alpha(x)^2 - f_\alpha(x)^2\Big] \,\mathrm{d}x \\
&= \frac{1}{2\sigma_\alpha^2}\int p_\alpha(x)\Big[f_\alpha(x)^2 - 2f_\alpha(x)f_\beta(x) + f_\beta(x)^2\Big] \,\mathrm{d}x \\
&= \frac{1}{2\sigma_\alpha^2}\int p_\alpha(x)\,\big(f_\alpha(x) - f_\beta(x)\big)^2 \,\mathrm{d}x. \qquad \text{(B.4)}
\end{aligned}
$$

B.2 Proof of Equation 6.11

Before proving Equation 6.11 from the main text, two inequalities must be stated. First, for non-negative numbers $k_1, k_2, ..., k_n$ and $l_1, l_2, ..., l_n$ the log-sum inequality is given by

$$\left(\sum_{i=1}^{n} k_i\right) \ln \frac{\sum_{i=1}^{n} k_i}{\sum_{i=1}^{n} l_i} \leq \sum_{i=1}^{n} k_i \ln \frac{k_i}{l_i}, \tag{B.5}$$

with equality if and only if k_i/l_i is equal to a constant value[145].

For the second inequality, we utilise the chain rule for relative entropy[145], given by

$$D\big(p_\alpha(x)||p_\beta(x)\big) + D\big(p_\alpha(y|x)||p_\beta(y|x)\big) = D\big(p_\alpha(x,y)||p_\beta(x,y)\big) \tag{B.6}$$

The non-negativity of $D\big(p_\alpha(x)||p_\beta(x)\big)$ implies that

$$D\big(p_\alpha(y|x)||p_\beta(y|x)\big) \leq D\big(p_\alpha(x,y)||p_\beta(x,y)\big). \tag{B.7}$$

Using the above inequalities,

$$
\begin{aligned}
D\big(p_\alpha(y|x)||p_\beta(y|x)\big) &\leq \iint p_\alpha(x,y)\,\ln\frac{p_\alpha(x,y)}{p_\beta(x,y)}\,\mathrm{d}x\,\mathrm{d}y \\
&= \iint\left(\sum_\theta p(x,y|\theta)\alpha(\theta)\right)\ln\frac{\sum_\theta p(x,y|\theta)\alpha(\theta)}{\sum_\theta p(x,y|\theta)\beta(\theta)}\,\mathrm{d}x\,\mathrm{d}y \\
&\leq \iint\sum_\theta p(x,y|\theta)\alpha(\theta)\,\ln\frac{p(x,y|\theta)\alpha(\theta)}{p(x,y|\theta)\beta(\theta)}\,\mathrm{d}x\,\mathrm{d}y \\
&= \iint\sum_\theta p(x,y|\theta)\alpha(\theta)\,\ln\frac{\alpha(\eta)}{\beta(\eta)}\,\mathrm{d}x\,\mathrm{d}y \\
&= \sum_\theta\left[\iint p(x,y|\theta)\,\mathrm{d}x\,\mathrm{d}y\right]\alpha(\theta)\,\ln\frac{\alpha(\eta)}{\beta(\eta)} \\
&= \sum_\theta \alpha(\eta)\,\ln\frac{\alpha(\eta)}{\beta(\eta)} \\
&= D\big(\alpha(\eta)||\beta(\eta)\big), \qquad \text{(B.8)}
\end{aligned}
$$

where the last line follows from the discrete definition of relative entropy using natural logarithms[145].

B.3 Proof of Equation 6.13

We define $p_\omega(x,y) = \sum_i^M \omega_i\, p_i(x,y)$: the average over M different complex selection probabilities. Following this definition, we also have $p_\omega(y|x) = \sum_i^M \omega_i p_i(x,y)/\sum_i^M \omega_i p_i(x)$. To prove Equation 11, we utilize a standard information theoretic manipulation[145] in the second line. Starting from the left hand side of Equation 11 we have:

$$
\begin{aligned}
\sum_i^M \omega_i \, D\big(p_i(y|x)||q(y|x)\big) &= \sum_i^M \omega_i \iint p_i(x,y) \, \ln \frac{p_i(y|x)}{q(y|x)} \, \mathrm{d}x \, \mathrm{d}y \\
&= \sum_i^M \omega_i \iint p_i(x,y) \, \ln \frac{p_i(y|x)}{q(y|x)} \frac{p_\omega(y|x)}{p_\omega(y|x)} \, \mathrm{d}x \, \mathrm{d}y \\
&= \sum_i^M \omega_i \iint p_i(x,y) \, \left[\ln \frac{p_i(y|x)}{p_\omega(y|x)} + \ln \frac{p_\omega(y|x)}{q(y|x)} \right] \mathrm{d}x \, \mathrm{d}y \\
&= \sum_i^M \omega_i \, D\big(p_i(y|x)||p_\omega(y|x)\big) + \sum_i^M \omega_i \iint p_i(x,y) \, \ln \frac{p_\omega(y|x)}{q(y|x)} \, \mathrm{d}x \, \mathrm{d}y \\
&= \sum_i^M \omega_i \, D\big(p_i(y|x)||p_\omega(y|x)\big) + \iint \left[\sum_i^M \omega_i \, p_i(x,y) \right] \ln \frac{p_\omega(y|x)}{q(y|x)} \, \mathrm{d}x \, \mathrm{d}y \\
&= \sum_i^M \omega_i \, D\big(p_i(y|x)||p_\omega(y|x)\big) + \iint p_\omega(x,y) \, \ln \frac{p_\omega(y|x)}{q(y|x)} \, \mathrm{d}x \, \mathrm{d}y \\
&= \sum_i^M \omega_i \, D\big(p_i(y|x)||p_\omega(y|x)\big) + D\big(p_\omega(y|x)||q(y|x)\big) \\
&\geq \sum_i^M \omega_i \, D\big(p_i(y|x)||p_\omega(y|x)\big),
\end{aligned}
$$

where the last line follows from the fact that $D\big(p_\omega(y|x)||q(y|x)\big) \geq 0$.

B.4 Proof of Equation 6.14

To prove Equation 12, first we show that the average cross entropy of the average density, $p_\omega(y|x)$, on each protein specific density, denoted $C_i(Y|X)$, is equal to the conditional entropy, $h_\omega(Y|X)$, of the average density:

$$\begin{aligned}\sum_i^M \omega_i\, C_i(Y|X) &= -\sum_i^M \omega_i \iint p_i(x,y)\, \ln p_\omega(y|x)\, \mathrm{d}x\, \mathrm{d}y \\ &= -\iint \left[\sum_i^M \omega_i\, p_i(x,y)\right] \ln p_\omega(y|x)\, \mathrm{d}x\, \mathrm{d}y \\ &= -\iint p_\omega(x,y)\, \ln p_\omega(y|x)\, \mathrm{d}x\, \mathrm{d}y \\ &= h_\omega(Y|X). \end{aligned} \tag{B.9}$$

In addition, from Equation 6.4, the average cross entropy

$$\begin{aligned}\sum_i^M \omega_i\, C_i(Y|X) &= \sum_i^M \omega_i \big[\, h_i(Y|X) + D\big(p_i(y|x)||p_\omega(y|x)\big)\,\big] \\ &= \sum_i^M \omega_i\, h_i(Y|X) + \sum_i^M \omega_i\, D\big(p_i(y|x)||p_\omega(y|x)\big). \end{aligned} \tag{B.10}$$

The inequality stated in Equation 6.14 arises by considering the second term in Equation B.10: the average relative entropy between a protein-specific density and the generalized density. As the relative entropy is non-negative, the average can be zero if and only if $D\big(p_i(y|x)||p_\omega(y|x)\big) = 0$ for all $i = 1, 2, ..., M$. However, each $p_i(y|x)$ is distinct. This means that at most only one i out of M can have $D\big(p_i(y|x)||p_\omega(y|x)\big) = 0$. The rest must have relative entropies greater than zero, so

$$\sum_i^M \omega_i\, D\big(p_i(y|x)||p_\omega(y|x)\big) > 0. \tag{B.11}$$

Combining Equations B.9, B.10 and B.11 we find

$$h_\omega(Y|X) = \sum_i^M \omega_i \, h_i(Y|X) + \sum_i^M \omega_i \, D\big(p_i(y|x)||p_\omega(y|x)\big)$$
$$h_\omega(Y|X) > \sum_i^M \omega_i \, h_i(Y|X),$$

as stated in the main text.

B.5 Proof of Equation 6.16

The proof of Equation 6.16 has two parts. Firstly, we derive a lower bound for $D\big(p(y|x,z)||q(y|x)\big)$, and secondly, re-express that lower bound in terms of mutual information. Using $p(y,x) = \int p(y,x,z)\ \mathrm{d}z$ and $p(y|x) = p(y,x)/p(x)$, we have

$$\begin{aligned}
D\big(p(y|x,z)||q(y|x)\big) &= \iiint p(y,x,z)\ \ln\frac{p(y|x,z)}{q(y|x)}\ \mathrm{d}y\ \mathrm{d}x\ \mathrm{d}z \\
&= \iiint p(y,x,z)\ \ln\frac{p(y|x,z)}{q(y|x)}\frac{p(y|x)}{p(y|x)}\ \mathrm{d}y\ \mathrm{d}x\ \mathrm{d}z \\
&= \iiint p(y,x,z)\left[\ln\frac{p(y|x,z)}{p(y|x)} + \ln\frac{p(y|x)}{q(y|x)}\right]\ \mathrm{d}y\ \mathrm{d}x\ \mathrm{d}z \\
&= D\big(p(y|x,z)||p(y|x)\big) + \iiint p(y,x,z)\ \ln\frac{p(y|x)}{q(y|x)}\ \mathrm{d}y\ \mathrm{d}x\ \mathrm{d}z \\
&= D\big(p(y|x,z)||p(y|x)\big)z + \iint\left[\int p(y,x,z)\ \mathrm{d}z\right]\ \ln\frac{p(y|x)}{q(y|x)}\ \mathrm{d}y\ \mathrm{d}x \\
&= D\big(p(y|x,z)||p(y|x)\big) + \iint p(y,x)\ \ln\frac{p(y|x)}{q(y|x)}\ \mathrm{d}y\ \mathrm{d}x \\
&= D\big(p(y|x,z)||p(y|x)\big) + D\big(p(y|x)||q(y|x)\big) \\
&\geq D\big(p(y|x,z)||p(y|x)\big), \qquad \text{(B.12)}
\end{aligned}$$

where the last line follows from the non-negativity of $D(p(y|x)||q(y|x))$. As the above is true for all $q(y|x)$, $p(y|x)$ is the scoring model that minimizes the regret in the case of missing information. The practical implications of this lower bound can be seen by noting that

$$\begin{aligned} D\big(p(y|x,z)||p(y|x)\big) &= \iiint p(y,x,z)\,\ln\frac{p(y|x,z)}{p(y|x)}\;\mathrm{d}y\;\mathrm{d}x\;\mathrm{d}z \\ &= \iiint p(y,x,z)\,\ln\frac{p(y|x,z)}{p(y|x)}\frac{p(z|x)}{p(z|x)}\;\mathrm{d}y\;\mathrm{d}x\;\mathrm{d}z \\ &= \iiint p(y,x,z)\,\ln\frac{p(y,z|x)}{p(y|x)p(z|x)}\;\mathrm{d}y\;\mathrm{d}x\;\mathrm{d}z \\ &= I(Y;Z|X), \end{aligned} \tag{B.13}$$

which is the mutual information of Y and Z given X. Using the chain rule of mutual information[145], this can be expressed as

$$I(Y;Z|X) = I(Y,X;Z) - I(X;Z). \tag{B.14}$$

The above represents the expected error for an optimal scoring function that has discarded data from Z. Thus, we have established that

$$D\big(p(y|x,z)||q(y|x)\big) \geq I(Y,X;Z) - I(X;Z). \tag{B.15}$$

Bibliography

1. S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nat. Rev. Drug Discov.*, vol. 9, pp. 203–214, Feb. 2010.

2. I. Kola and J. Landis, "Can the pharmaceutical industry reduce attrition rates?," *Nat. Rev. Drug Discov.*, vol. 3, pp. 711–716, Aug. 2004.

3. A. Mullard, "2010 FDA drug approvals," *Nat. Rev. Drug Discov.*, vol. 10, pp. 82–85, Feb. 2011.

4. W. L. Jorgensen, "The Many Roles of Computation in Drug Discovery," *Science*, vol. 303, pp. 1813–1818, Mar. 2004.

5. J. Drie, "Computer-aided drug design: the next 20 years," *J. Comput.-Aided Mol. Des.*, vol. 21, pp. 591–601, Oct. 2007.

6. G. Schneider, "Virtual screening: an endless staircase?," *Nat. Rev. Drug Discov.*, vol. 9, pp. 273–276, Apr. 2010.

7. D. Green, A. Leach, and M. Head, "Computer-aided molecular design under the SWOT-light," *J. Comput.-Aided Mol. Des.*, vol. 26, pp. 51–56, Dec. 2012.

8. M. Dickson and J. P. Gagnon, "Key factors in the rising cost of new drug discovery and development," *Nat. Rev. Drug Discov.*, vol. 3, pp. 417–429, May 2004.

9. D. A. Pereira and J. A. Williams, "Origin and evolution of high throughput screening," *Br. J. Pharmacol.*, vol. 152, pp. 53–61, Sept. 2007.

10. C. Bergsdorf and J. Ottl, "Affinity-based screening techniques: their impact and benefit to increase the number of high quality leads.," *Expert Opin. Drug Discov.*, vol. 5, pp. 1095–1107, Nov. 2010.

11. E. A. Martis, Radhakrishnan, and Badve, "High-Throughput Screening: The Hits and Leads of Drug Discovery - An Overview," *J. App. Pharm. Science*, vol. 1, no. 1, pp. 2–10, 2011.

12. H. Köppen, "Virtual screening - what does it give us?," *Curr. Opin. Drug Discovery Dev.*, vol. 12, pp. 397–407, May 2009.

13. K. H. Bleicher, H.-J. Bohm, K. Muller, and A. I. Alanine, "Hit and lead generation: beyond high-throughput screening," *Nat. Rev. Drug Discov.*, vol. 2, pp. 369–378, May 2003.

14. G. Klebe, "Recent developments in structure-based drug design," *J. Mol. Med.*, vol. 78, pp. 269–281, July 2000.

15. C. Bissantz, B. Kuhn, and M. Stahl, "A medicinal chemist's guide to molecular interactions.," *J. Med. Chem.*, vol. 53, pp. 5061–5084, July 2010.

16. A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain, and B. Kelley, "Molecular Shape and Medicinal Chemistry: A Perspective," *J. Med. Chem.*, vol. 53, pp. 3862–3886, Feb. 2010.

17. D. J. Huggins, W. Sherman, and B. Tidor, "Rational Approaches to Improving Selectivity in Drug Design," *J. Med. Chem.*, vol. 55, pp. 1424–1444, Jan. 2012.

18. H. Gohlke and G. Klebe, "Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors.," *Angew. Chem. Int. Ed.*, vol. 41, pp. 2644–2676, Aug. 2002.

19. N. Nikolova and J. Jaworska, "Approaches to measure chemical similarity a review," *QSAR Comb. Sci.*, vol. 22, pp. 1006–1026, Dec. 2003.

20. P. Willett, J. M. Barnard, and G. M. Downs, "Chemical Similarity Searching," *J. Chem. Inf. Comput. Sci.*, vol. 38, pp. 983–996, July 1998.

21. P. J. Ballester and W. G. Richards, "Ultrafast shape recognition to search compound databases for similar molecular shapes.," *J. Comput. Chem.*, vol. 28, pp. 1711–1723, July 2007.

22. M. S. Armstrong, G. M. Morris, P. W. Finn, R. Sharma, L. Moretti, R. I. Cooper, and W. G. Richards, "ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics.," *J. Comput.-Aided Mol. Des.*, vol. 24, pp. 789–801, Sept. 2010.

23. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, "A geometric approach to macromolecule-ligand interactions.," *J. Mol. Biol.*, vol. 161, pp. 269–288, Oct. 1982.

24. T. Cheng, Q. Li, Z. Zhou, Y. Wang, and S. H. Bryant, "Structure-based virtual screening for drug discovery: a problem-centric review.," *AAPS J.*, vol. 14, pp. 133–141, Mar. 2012.

25. G. L. Warren, W. W. Andrews, A.-M. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head, "A critical assessment of docking programs and scoring functions.," *J. Med. Chem.*, vol. 49, pp. 5912–5931, Oct. 2006.

26. H. Chen, P. D. Lyne, F. Giordanetto, T. Lovell, and J. Li, "On evaluating molecular-docking methods for pose prediction and enrichment factors.," *J. Chem. Inf. Model.*, vol. 46, no. 1, pp. 401–415, 2006.

27. P. Ferrara, H. Gohlke, D. J. Price, G. Klebe, and C. L. Brooks, "Assessing scoring functions for protein-ligand interactions.," *J. Med. Chem.*, vol. 47, pp. 3032–3047, June 2004.

28. T. Cheng, X. Li, Y. Li, Z. Liu, and R. Wang, "Comparative Assessment of Scoring Functions on a Diverse Test Set," *J. Chem. Inf. Model.*, vol. 49, pp. 1079–1093, Apr. 2009.

29. I. J. Enyedy and W. J. Egan, "Can we use docking and scoring for hit-to-lead optimization?," *J. Comput.-Aided Mol. Des.*, vol. 22, pp. 161–168, Mar. 2008.

30. M. L. Verdonk, V. Berdini, M. J. Hartshorn, W. T. M. Mooij, C. W. Murray, R. D. Taylor, and P. Watson, "Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment," *J. Chem. Inf. Comput. Sci.*, vol. 44, pp. 793–806, Feb. 2004.

31. N. Huang, B. K. Shoichet, and J. J. Irwin, "Benchmarking Sets for Molecular Docking," *J. Med. Chem.*, vol. 49, pp. 6789–6801, Oct. 2006.

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

33. J. D. Durrant, A. J. Friedman, K. E. Rogers, and J. A. McCammon, "Comparing Neural-Network Scoring Functions and the State of the Art: Applications to Common Library Screening," *J. Chem. Inf. Model.*, vol. 53, pp. 1726–1735, June 2013.

34. A. M. Ruvinsky and A. V. Kozintsev, "New and fast statistical-thermodynamic method for computation of protein-ligand binding entropy substantially improves docking accuracy.," *J. Comput. Chem.*, vol. 26, pp. 1089–1095, Aug. 2005.

35. J. Lee and C. Seok, "A statistical rescoring scheme for proteinligand docking: Consideration of entropic effect," *Proteins: Struct., Funct., Bioinf.*, vol. 70, no. 3, pp. 1074–1083, 2008.

36. A. V. Grigoryan, H. Wang, and T. J. Cardozo, "Can the Energy Gap in the Protein-Ligand Binding Energy Landscape Be Used as a Descriptor in Virtual Ligand Screening?," *PLoS ONE*, vol. 7, pp. e46532+, Oct. 2012.

37. J.-H. H. Lin, A. L. Perryman, J. R. Schames, and J. A. McCammon, "The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme.," *Biopolymers*, vol. 68, pp. 47–62, Jan. 2003.

38. T. Hou, J. Wang, Y. Li, and W. Wang, "Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations," *J. Chem. Inf. Model.*, vol. 51, pp. 69–82, Nov. 2010.

39. J. Åqvist, C. Medina, and J.-E. Samuelsson, "A new method for predicting binding affinity in computer-aided drug design," *Protein Eng.*, vol. 7, pp. 385–391, Mar. 1994.

40. M. Shirts, D. Mobley, and J. Chodera, "Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time?," *Annu. Rep. Med. Chem.*, vol. 3, pp. 41–59, 2007.

41. J. Michel and J. Essex, "Prediction of proteinligand binding affinity by free energy simulations: assumptions, pitfalls and expectations," *J. Comput.-Aided Mol. Des.*, vol. 24, pp. 639–658, Aug. 2010.

42. W. L. Jorgensen, M. Bollini, V. V. Thakur, R. A. Domaoal, K. A. Spasov, and K. S. Anderson, "Efficient Discovery of Potent Anti-HIV Agents Targeting the Tyr181Cys Variant of HIV Reverse Transcriptase," *J. Am. Chem. Soc.*, vol. 133, pp. 15686–15696, Aug. 2011.

43. J. C. Gumbart, B. Roux, and C. Chipot, "Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy?," *J. Chem. Theory Comput.*, vol. 9, pp. 794–802, Nov. 2012.

44. J. Michel and J. W. Essex, "Hit identification and binding mode predictions by rigorous free energy simulations.," *J. Med. Chem.*, vol. 51, pp. 6654–6664, Nov. 2008.

45. R. D. Malmstrom and S. J. Watowich, "Using Free Energy of Binding Calculations To Improve the Accuracy of Virtual Screening Predictions," *J. Chem. Inf. Model.*, vol. 51, pp. 1648–1655, June 2011.

46. J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande, "Alchemical free energy methods for drug discovery: progress and challenges.," *Curr. Opin. Struc. Biol.*, vol. 21, pp. 150–160, Apr. 2011.

47. R. D. Smith, J. B. Dunbar, P. M. Ung, E. X. Esposito, C.-Y. Yang, S. Wang, and H. A. Carlson, "CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions," *J. Chem. Inf. Model.*, vol. 51, pp. 2115–2131, Aug. 2011.

48. P. Ball, "Water as an active constituent in cell biology.," *Chem. Rev.*, vol. 108, pp. 74–108, Jan. 2008.

49. H. G. Wallnoefer, S. Handschuh, K. R. Liedl, and T. Fox, "Stabilizing of a globular protein by a highly complex water network: a molecular dynamics simulation study on factor Xa.," *J. Phys. Chem. B*, vol. 114, pp. 7405–7412, June 2010.

50. C. D. Dellisanti, S. M. Hanson, L. Chen, and C. Czajkowski, "Packing of the extracellular domain hydrophobic core has evolved to facilitate pentameric ligand-gated ion channel function.," *J. Biol. Chem.*, vol. 286, pp. 3658–3670, Feb. 2011.

51. O. Beckstein and M. S. P. Sansom, "Liquidvapor oscillations of water in hydrophobic nanopores," *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 7063–7068, June 2003.

52. J. Ostmeyer, S. Chakrapani, A. C. Pan, E. Perozo, and B. Roux, "Recovery from slow inactivation in K+ channels is controlled by water molecules," *Nature*, vol. 501, pp. 121–124, Sept. 2013.

53. Y. Lu, R. Wang, C.-Y. Yang, and S. Wang, "Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structures of Protein-Ligand Complexes," *J. Chem. Inf. Model.*, vol. 47, pp. 668–675, Feb. 2007.

54. S. H. Sleigh, P. R. Seavers, A. J. Wilkinson, J. E. Ladbury, and J. R. H. Tame, "Crystallographic and Calorimetric Analysis of Peptide Binding to OppA Protein," *J. Mol. Biol.*, vol. 291, pp. 393–415, Aug. 1999.

55. P. Cozzini, M. Fornabaio, A. Marabotti, D. J. Abraham, G. E. Kellogg, and A. Mozzarelli, "Free Energy of Ligand Binding to Protein: Evaluation of the Contribution of Water Molecules by Computational Methods," *Curr. Med. Chem.*, pp. 3093–3118, Dec. 2004.

56. Z. Li and T. Lazaridis, "Water at biomolecular binding interfaces.," *Phys. Chem. Chem. Phys.*, vol. 9, pp. 573–581, Feb. 2007.

57. S. B. de Beer, N. P. Vermeulen, and C. Oostenbrink, "The role of water molecules in computational drug design.," *Curr. Top. Med. Chem.*, vol. 10, no. 1, pp. 55–66, 2010.

58. M. Pastor, G. Cruciani, and K. A. Watson, "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.*, vol. 40, pp. 4089–4102, Dec. 1997.

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

60. A. Hussain, J. Melville, and J. Hirst, "Molecular docking and QSAR of aplyronine A and analogues: potent inhibitors of actin," *J. Comput.-Aided Mol. Des.*, vol. 24, pp. 1–15, Jan. 2010.

61. M. O. Taha, M. Habash, Z. Al-Hadidi, A. Al-Bakri, K. Younis, and S. Sisan, "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.*, vol. 51, pp. 647–669, Mar. 2011.

62. J. Luccarelli, J. Michel, J. Tirado-Rives, and W. L. Jorgensen, "Effects of Water Placement on Predictions of Binding Affinities for p38a MAP Kinase Inhibitors," *J. Chem. Theory Comput.*, vol. 6, pp. 3850–3856, Dec. 2010.

63. A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski, "Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures.," *FEBS J.*, vol. 275, pp. 1–21, Jan. 2008.

64. H. S. Frank and M. W. Evans, "Free Volume and Entropy in Condensed Systems III. Entropy in Binary Liquid Mixtures; Partial Molal Entropy in Dilute Solutions; Structure and Thermodynamics in Aqueous Electrolytes," *J. Chem. Phys.*, vol. 13, no. 11, pp. 507–532, 1945.

65. W. Kauzmann, *Some Factors in the Interpretation of Protein Denaturation*, vol. 14, pp. 1–63. Adv. Protein Chem., 1959.

66. C. Tanford, "Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins," *J. Am. Chem. Soc.*, vol. 84, pp. 4240–4247, Nov. 1962.

67. W. Blokzijl and J. B. F. N. Engberts, "Hydrophobic Effects. Opinions and Facts," *Angew. Chem., Int. Ed.*, vol. 32, no. 11, pp. 1545–1579, 1993.

68. P. D. Ross and S. Subramanian, "Thermodynamics of protein association reactions: forces contributing to stability," *Biochemistry*, vol. 20, pp. 3096–3102, May 1981.

69. L. R. Pratt and D. Chandler, "Theory of the hydrophobic effect," *J. Chem. Phys.*, vol. 67, no. 8, pp. 3683–3704, 1977.

70. G. Hummer, S. Garde, A. E. García, A. Pohorille, and L. R. Pratt, "An information theory model of hydrophobic interactions," *Proc. Natl. Acad. Sci. USA*, vol. 93, pp. 8951–8955, Aug. 1996.

71. K. Lum, D. Chandler, and J. D. Weeks, "Hydrophobicity at Small and Large Length Scales," *J. Phys. Chem. B*, vol. 103, pp. 4570–4577, Apr. 1999.

72. T. M. Raschke, J. Tsai, and M. Levitt, "Quantification of the hydrophobic interaction by simulations of the aggregation of small hydrophobic solutes in water.," *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 5965–5969, May 2001.

73. R. Zangi, "Driving Force for Hydrophobic Interaction at Different Length Scales," *J. Phys. Chem. B*, vol. 115, pp. 2303–2311, Mar. 2011.

74. C. Y. Lee, J. A. McCammon, and P. J. Rossky, "The structure of liquid water at an extended hydrophobic surface," *J. Chem. Phys.*, vol. 80, no. 9, pp. 4448–4455, 1984.

75. R. Steitz, T. Gutberlet, T. Hauss, B. Klösgen, R. Krastev, S. Schemmel, A. C. Simonsen, and G. H. Findenegg, "Nanobubbles and Their Precursor Layer at the Interface of Water Against a Hydrophobic Substrate," *Langmuir*, vol. 19, pp. 2409–2418, Feb. 2003.

76. I. Brovchenko, D. Paschek, and A. Geiger, "Gibbs ensemble simulation of water in spherical cavities," *J. Chem. Phys.*, vol. 113, no. 12, pp. 5026–5036, 2000.

77. P. Setny, R. Baron, and J. A. McCammon, "How Can Hydrophobic Association Be Enthalpy Driven?," *J. Chem. Theory Comput.*, vol. 6, pp. 2866–2871, Sept. 2010.

78. D. Chandler, "Interfaces and the driving force of hydrophobic assembly," *Nature*, vol. 437, pp. 640–647, Sept. 2005.

79. K. A. Dill, T. M. Truskett, V. Vlachy, and B. Hribar-Lee, "Modeling water, the hydrophobic effect, and ion solvation.," *Annu. Rev. Bioph. Biom.*, vol. 34, no. 1, pp. 173–199, 2005.

80. K. A. Dill and S. Bromberg, *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience, 2nd Edition.* Garland Science, 2nd ed., Dec. 2010.

81. T. P. Silverstein, "The Real Reason Why Oil and Water Don't Mix," *J. Chem. Educ.*, vol. 75, pp. 116+, Jan. 1998.

82. A. Ben-Naim, "The Rise and Fall of the Hydrophobic Effect in Protein Folding and Protein-Protein Association, and Molecular Recognition," *Open J. Biophys.*, vol. 1, pp. 1–7, 2011.

83. G. Graziano, "A view on the dogma of hydrophobic imperialism in protein folding," *J. Biomol. Struct. Dyn.*, vol. 31, pp. 1016–1019, Jan. 2013.

84. J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry, Fifth Edition: International Version (hardcover).* W. H. Freeman, fifth edition ed., Feb. 2002.

85. L. A. Moran, R. A. Horton, G. Scrimgeour, and M. Perry, *Principles of Biochemistry (5th Edition).* Prentice Hall, 5 ed., Sept. 2011.

86. M. Daune, *Molecular Biophysics: Structures in Motion.* Oxford University Press, USA, Apr. 1999.

87. D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry, Fourth Edition.* W. H. Freeman, 4th ed., Apr. 2004.

88. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.," *Adv. Drug Deliver. Rev.*, vol. 46, pp. 3–26, Mar. 2001.

89. J. D. Dunitz, "The entropic cost of bound water in crystals and biomolecules.," *Science*, vol. 264, p. 670, Apr. 1994.

90. P. Y. Lam, P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, and A. Et, "Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors," *Science*, vol. 263, pp. 380–384, Jan. 1994.

91. J. M. Chen, S. L. Xu, Z. Wawrzak, G. S. Basarab, and D. B. Jordan, "Structure-Based Design of Potent Inhibitors of Scytalone Dehydratase: Displacement of a Water Molecule from the Active Site," *Biochemistry*, vol. 37, pp. 17735–17744, Dec. 1998.

92. A. Wissner, D. M. Berger, D. H. Boschelli, M. B. Floyd, L. M. Greenberger, B. C. Gruber, B. D. Johnson, N. Mamuya, R. Nilakantan, M. F. Reich, R. Shen, H.-R. Tsou, E. Upeslacis, Y. F. Wang, B. Wu, F. Ye, and N. Zhang, "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.*, vol. 43, pp. 3244–3256, Aug. 2000.

93. C. Clarke, R. J. Woods, J. Gluska, A. Cooper, M. A. Nutley, and G.-J. Boons, "Involvement of Water in Carbohydrate-Protein Binding," *J. Am. Chem. Soc.*, vol. 123, pp. 12238–12247, Dec. 2001.

94. C. Liu, S. T. Wrobleski, J. Lin, G. Ahmed, A. Metzger, J. Wityak, K. M. Gillooly, D. J. Shuster, K. W. McIntyre, S. Pitt, D. R. Shen, R. F. Zhang, H. Zhang, A. M. Doweyko, D. Diller, I. Henderson, J. C. Barrish, J. H. Dodd, G. L. Schieven, and K. Leftheris, "5-Cyanopyrimidine Derivatives as a Novel Class of Potent, Selective, and Orally Active Inhibitors of p38a MAP Kinase," *J. Med. Chem.*, vol. 48, pp. 6261–6270, Oct. 2005.

95. R. Kadirvelraj, B. L. Foley, J. D. Dyekjær, and R. J. Woods, "Involvement of Water in Carbohydrate-Protein Binding: Concanavalin A Revisited," *J. Am. Chem. Soc.*, vol. 130, pp. 16933–16942, Nov. 2008.

96. V. Mikol, C. Papageorgiou, and X. Borer, "The Role of Water Molecules in the Structure-Based Design of (5-Hydroxynorvaline)-2-cyclosporin: Synthesis, Biological Activity, and Crystallographic Analysis with Cyclophilin A," *J. Med. Chem.*, vol. 38, pp. 3361–3367, Aug. 1995.

97. N. N. Nasief, H. Tan, J. Kong, and D. Hangauer, "Water Mediated Ligand Functional Group Cooperativity: The Contribution of a Methyl Group to Binding Affinity is Enhanced by a COO Group Through Changes in the Structure and Thermodynamics of the Hydration Waters of LigandThermolysin Complexes," *J. Med. Chem.*, vol. 55, pp. 8283–8302, Aug. 2012.

98. F. Vollmuth and M. Geyer, "Interaction of Propionylated and Butyrylated Histone H3 Lysine Marks with Brd4 Bromodomains," *Angew. Chem. Int. Ed.*, vol. 49, no. 38, pp. 6768–6772, 2010.

99. A. Biela, F. Sielaff, F. Terwesten, A. Heine, T. Steinmetzer, and G. Klebe, "Ligand Binding Stepwise Disrupts Water Network in Thrombin: Enthalpic and Entropic Changes Reveal Classical Hydrophobic Effect," *J. Med. Chem.*, May 2012.

100. P. W. Snyder, J. Mecinović, D. T. Moustakas, S. W. Thomas, M. Harder, E. T. Mack, M. R. Lockett, A. Héroux, W. Sherman, and G. M. Whitesides, "Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase," *Proc. Natl. Acad. Sci. USA*, vol. 108, pp. 17889–17894, Nov. 2011.

101. J. Mecinović, P. W. Snyder, K. A. Mirica, S. Bai, E. T. Mack, R. L. Kwant, D. T. Moustakas, A. Héroux, and G. M. Whitesides, "Fluoroalkyl and Alkyl Chains Have Similar Hydrophobicities in Binding to the "Hydrophobic Wall" of Carbonic Anhydrase," *J. Am. Chem. Soc.*, vol. 133, pp. 14017–14026, July 2011.

102. J. M. Myslinski, J. E. DeLorbe, J. H. Clements, and S. F. Martin, "Protein-ligand interactions: thermodynamic effects associated with increasing nonpolar surface area.," *J. Am. Chem. Soc.*, vol. 133, pp. 18518–18521, Nov. 2011.

103. L. Englert, A. Biela, M. Zayed, A. Heine, D. Hangauer, and G. Klebe, "Displacement of disordered water molecules from hydrophobic pocket creates enthalpic signature: Binding of phosphonamidate to the S1'-pocket of thermolysin," *Biochim. Biophys. Acta.*, vol. 1800, pp. 1192–1202, Nov. 2010.

104. R. Talhout, A. Villa, A. E. Mark, and J. B. F. N. Engberts, "Understanding Binding Affinity: A Combined Isothermal Titration Calorimetry/Molecular Dynamics Study of the Binding of a Series of Hydrophobically Modified Benzamidinium Chloride Inhibitors to Trypsin," *J. Am. Chem. Soc.*, vol. 125, pp. 10570–10579, Aug. 2003.

105. A. T. Fenley, H. S. Muddana, and M. K. Gilson, "Entropyenthalpy transduction caused by conformational shifts can obscure the forces driving proteinligand binding," *Proc. Natl. Acad. Sci. USA*, vol. 109, pp. 20006–20011, Dec. 2012.

106. B. Yu, M. Blaber, A. M. Gronenborn, G. M. Clore, and D. L. D. Caspar, "Disordered water within a hydrophobic protein cavity visualized by x-ray crystallography," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 103–108, Jan. 1999.

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

108. O. Beckstein, P. C. Biggin, and M. S. P. Sansom, "A Hydrophobic Gating Mechanism for Nanopores," *J. Phys. Chem. B*, vol. 105, pp. 12902–12905, Dec. 2001.

109. J. Qvist, M. Davidovic, D. Hamelberg, and B. Halle, "A dry ligand-binding cavity in a solvated protein," *Proc. Natl. Acad. Sci. USA*, vol. 105, pp. 6296–6301, Apr. 2008.

110. R. J. Bingham, J. B. Findlay, S.-Y. Y. Hsieh, A. P. Kalverda, A. Kjellberg, C. Perazzolo, S. E. Phillips, K. Seshadri, C. H. Trinh, W. B. Turnbull, G. Bodenhausen, and S. W. Homans, "Thermodynamics of binding of 2-methoxy-3-isopropylpyrazine and 2-methoxy-3-isobutylpyrazine to the major urinary protein.," *J. Am. Chem. Soc.*, vol. 126, pp. 1675–1681, Feb. 2004.

111. E. Barratt, R. J. Bingham, D. J. Warner, C. A. Laughton, S. E. V. Phillips, and S. W. Homans, "Van der Waals Interactions Dominate Ligand-Protein Association in a Protein Binding Site Occluded from Solvent Water," *J. Am. Chem. Soc.*, vol. 127, pp. 11827–11834, July 2005.

112. A. T. García-Sosa, "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.*, May 2013.

113. A. Biela, N. N. Nasief, M. Betz, A. Heine, D. Hangauer, and G. Klebe, "Dissecting the Hydrophobic Effect on the Molecular Level: The Role of Water, Enthalpy, and Entropy in Ligand Binding to Thermolysin," *Angew. Chem., Int. Ed.*, vol. 52, pp. 1822–1828, Feb. 2013.

114. A. Bortolato, B. G. Tehan, M. S. Bodnarchuk, J. W. Essex, and J. S. Mason, "Water Network Perturbation in Ligand Binding: Adenosine A2A Antagonists as a Case Study," *J. Chem. Inf. Model.*, June 2013.

115. D. Hamelberg and J. A. McCammon, "Standard Free Energy of Releasing a Localized Water Molecule from the Binding Pockets of Proteins: Double-Decoupling Method," *J. Am. Chem. Soc.*, vol. 126, pp. 7683–7689, May 2004.

116. C. Barillari, J. Taylor, R. Viner, and J. W. Essex, "Classification of Water Molecules in Protein Binding Sites," *J. Am. Chem. Soc.*, vol. 129, pp. 2577–2587, Feb. 2007.

117. J. Michel, J. Tirado-Rives, and W. L. Jorgensen, "Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization," *J. Am. Chem. Soc.*, vol. 131, pp. 15403–15411, Sept. 2009.

118. A. T. García-Sosa and R. L. Mancera, "Free Energy Calculations of Mutations Involving a Tightly Bound Water Molecule and Ligand Substitutions in a Ligand-Protein Complex," *Mol. Inf.*, vol. 29, no. 8-9, pp. 589–600, 2010.

119. T. Lazaridis, "Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory," *J. Phys. Chem. B*, vol. 102, pp. 3531–3541, Apr. 1998.

120. T. Lazaridis, "Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids," *J. Phys. Chem. B*, vol. 102, pp. 3542–3550, Apr. 1998.

121. D. C. Wallace, "On the role of density fluctuations in the entropy of a fluid," *J. Chem. Phys.*, vol. 87, no. 4, pp. 2282–2284, 1987.

122. A. Baranyai and D. J. Evans, "Direct entropy calculation from computer simulation of liquids.," *Phys. Rev. A*, vol. 40, pp. 3817–3822, Oct. 1989.

123. T. Young, R. Abel, B. Kim, B. J. Berne, and R. A. Friesner, "Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding.," *Proc. Natl. Acad. Sci. USA*, vol. 104, pp. 808–813, Jan. 2007.

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

125. D. D. Robinson, W. Sherman, and R. Farid, "Understanding Kinase Selectivity Through Energetic Analysis of Binding Site Waters," *ChemMedChem*, vol. 5, pp. 618–627, Apr. 2010.

126. T. Beuming, R. Farid, and W. Sherman, "High-energy water sites determine peptide binding affinity and specificity of PDZ domains," *Protein Sci.*, vol. 18, pp. 1609–1619, Aug. 2009.

127. D. J. Huggins, M. Marsh, and M. C. Payne, "Thermodynamic Properties of Water Molecules at a ProteinProtein Interaction Surface," *J. Chem. Theory Comput.*, vol. 7, pp. 3514–3522, Sept. 2011.

128. D. J. Huggins, "Application of inhomogeneous fluid solvation theory to model the distribution and thermodynamics of water molecules around biomolecules," *Phys. Chem. Chem. Phys.*, vol. 14, no. 43, pp. 15106–15117, 2012.

129. Z. Li and T. Lazaridis, "Thermodynamic Contributions of the Ordered Water Molecule in HIV-1 Protease," *J. Am. Chem. Soc.*, vol. 125, pp. 6636–6637, May 2003.

130. L. Wang, B. J. Berne, and R. A. Friesner, "Ligand binding to protein-binding pockets with wet and dry regions," *Proc. Natl. Acad. Sci. USA*, vol. 108, pp. 1326–1330, Jan. 2011.

131. J. B. Dunbar, R. D. Smith, C.-Y. Yang, P. M. Ung, K. W. Lexa, N. A. Khazanov, J. A. Stuckey, S. Wang, and H. A. Carlson, "CSAR Benchmark Exercise of 2010: Selection of the ProteinLigand Complexes," *J. Chem. Inf. Model.*, vol. 51, pp. 2036–2046, July 2011.

132. G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 86, pp. 114–117, Apr. 1965.

133. H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Mol. Biol.*, vol. 10, p. 980, Dec. 2003.

134. R. Wang, X. Fang, Y. Lu, and S. Wang, "The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures," *J. Med. Chem.*, vol. 47, pp. 2977–2980, May 2004.

135. L. Hu, M. L. Benson, R. D. Smith, M. G. Lerner, and H. A. Carlson, "Binding MOAD (mother of all databases).," *Proteins: Struct., Funct., Bioinf.*, vol. 60, pp. 333–340, Aug. 2005.

136. T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.," *Nucleic Acids Res.*, vol. 35, pp. D198–D201, Jan. 2007.

137. J. B. Dunbar, R. D. Smith, K. L. Damm-Ganamet, A. Ahmed, E. X. X. Esposito, J. Delproposto, K. Chinnaswamy, Y.-N. N. Kang, G. Kubish, J. E. Gestwicki, J. A. Stuckey, and H. A. Carlson, "CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys.," *J. Chem. Inf. Model.*, vol. 53, pp. 1842–1852, Aug. 2013.

138. P. A. M. Dirac, *The Principles of Quantum Mechanics (International Series of Monographs on Physics)*. Oxford University Press, USA, 4 ed., Feb. 1982.

139. D. A. McQuarrie, *Statistical Mechanics*. University Science Books, 1st ed., June 2000.

140. E. Jaynes, "Information theory and statistical mechanics," *Phy. Rev.*, vol. 106, pp. 620–630, May 1957.

141. E. Jaynes, "Information theory and statistical mechanics. II," *Phys. Rev.*, vol. 108, pp. 171–190, Oct. 1957.

142. D. Frenkel and B. Smit, *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science)*. Academic Press, 2 ed., Nov. 2001.

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

144. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 3rd ed., Feb. 2009.

145. T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2nd ed., July 2006.

146. C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

147. K. P. Burnham and D. Anderson, *Model Selection and Multi-Model Inference*. Springer, 2nd ed., July 2002.

148. A. T. García-Sosa and R. L. Mancera, "The effect of a tightly bound water molecule on scaffold diversity in the computer-aided de novo ligand design of CDK2 inhibitors," *J. Mol. Model.*, vol. 12, pp. 422–431, Mar. 2006.

149. J. Günther, A. Bergner, M. Hendlich, and G. Klebe, "Utilising Structural Knowledge in Drug Design Strategies: Applications Using Relibase," *J. Mol. Biol.*, vol. 326, pp. 621–636, Feb. 2003.

150. O. Carugo and D. Bordo, "How many water molecules can be detected by protein crystallography?," *Acta Crystallogr. D*, vol. 55, pp. 479–483, Feb. 1999.

151. A. M. Davis, S. J. Teague, and G. J. Kleywegt, "Application and Limitations of X-ray Crystallographic Data in Structure-Based Ligand and Drug Design," *Angew. Chem. Int. Ed.*, vol. 42, pp. 2718–2736, June 2003.

152. N. V. Nucci, M. S. Pometun, and A. J. Wand, "Site-resolved measurement of water-protein interactions by solution NMR," *Nat. Struct. Mol. Biol.*, vol. 18, pp. 245–249, Feb. 2011.

153. R. H. Henchman and J. A. McCammon, "Extracting hydration sites around proteins from explicit water simulations.," *J. Comput. Chem.*, vol. 23, pp. 861–869, July 2002.

154. R. Vijayan, M. A. Sahai, T. Czajkowski, and P. C. Biggin, "A comparative analysis of the role of water in the binding pockets of ionotropic glutamate receptors.," *Phys. Chem. Chem. Phys.*, vol. 12, pp. 14057–14066, Nov. 2010.

155. H. Resat and M. Mezei, "Grand Canonical Monte Carlo Simulation of Water Positions in Crystal Hydrates," *J. Am. Chem. Soc.*, vol. 116, pp. 7451–7452, Aug. 1994.

156. J. Michel, J. Tirado-Rives, and W. L. Jorgensen, "Prediction of the Water Content in Protein Binding Sites," *J. Phys. Chem. B*, vol. 113, pp. 13337–13346, Sept. 2009.

157. Y. Liu and T. Ichiye, "An integral equation theory for the structure of water around globular solutes," *Chem. Phys. Lett.*, vol. 231, pp. 380–386, Dec. 1994.

158. T. Imai, R. Hiraoka, A. Kovalenko, and F. Hirata, "Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation," *Proteins: Struct., Funct., Bioinf.*, vol. 66, pp. 804–813, Mar. 2007.

159. K. Hirano, D. Yokogawa, H. Sato, and S. Sakaki, "An Analysis of 3D Solvation Structure in Biomolecules: Application to Coiled Coil Serine and Bacteriorhodopsin," *J. Phys. Chem. B*, vol. 114, pp. 7935–7941, June 2010.

160. D. J. Sindhikara, N. Yoshida, and F. Hirata, "Placevent: An algorithm for prediction of explicit solvent atom distribution-Application to HIV-1 protease and F-ATP synthase," *J. Comput. Chem.*, vol. 33, pp. 1536–1543, July 2012.

161. D. J. Sindhikara and F. Hirata, "Analysis of Biomolecular Solvation Sites by 3D-RISM Theory," *J. Phys. Chem. B*, vol. 117, pp. 6718–6723, May 2013.

162. P. J. Goodford, "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules," *J. Med. Chem.*, vol. 28, pp. 849–857, July 1985.

163. R. C. Wade and P. J. Goodford, "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.*, vol. 36, pp. 148–156, Jan. 1993.

164. P. Setny and M. Zacharias, "Hydration in Discrete Water. A Mean Field, Cellular Automata Based Approach to Calculating Hydration Free Energies," *J. Phys. Chem. B*, vol. 114, pp. 8667–8675, July 2010.

165. N. Thanki, J. Thornton, and J. Goodfellow, "Distributions of water around amino acid residues in proteins," *J. Mol. Biol.*, vol. 202, pp. 637–657, Aug. 1988.

166. W. R. Pitt and J. M. Goodfellow, "Modelling of solvent positions around polar groups in proteins," *Protein Eng.*, vol. 4, pp. 531–537, June 1991.

167. M. L. Verdonk, J. C. Cole, and R. Taylor, "SuperStar: a knowledge-based approach for identifying interaction sites in proteins.," *J Mol Biol*, vol. 289, pp. 1093–1108, June 1999.

168. J. W. Schymkowitz, F. Rousseau, I. C. Martins, J. Ferkinghoff-Borg, F. Stricher, and L. Serrano, "Prediction of water and metal binding sites and their affinities by using the Fold-X force field.," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 10147–10152, July 2005.

169. G. Rossato, B. Ernst, A. Vedani, and M. Smiesko, "AcquaAlta: a directional approach to the solvation of ligand-protein complexes.," *J. Chem. Inf. Model.*, vol. 51, pp. 1867–1881, Aug. 2011.

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

171. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235–242, Jan. 2000.

172. F. H. Allen, "The Cambridge Structural Database: a quarter of a million crystal structures and rising," *Acta Crystallogr. B*, vol. 58, pp. 380–388, June 2002.

173. H.-H. Bui, A. J. Schiewe, and I. S. Haworth, "WATGEN: An algorithm for modeling water networks at proteinprotein interfaces," *J. Comput. Chem.*, vol. 28, pp. 2241–2251, Nov. 2007.

174. D. J. Huggins and B. Tidor, "Systematic placement of structural water molecules for improved scoring of protein-ligand interactions," *Protein Eng. Des. Sel.*, vol. 24, pp. 777–789, July 2011.

175. S. Forli and A. J. Olson, "A Force Field with Discrete Displaceable Waters and Desolvation Entropy for Hydrated Ligand Docking," *J. Med. Chem.*, vol. 55, pp. 623–638, Dec. 2011.

176. O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.," *J. Comput. Chem.*, vol. 31, pp. 455–461, Jan. 2010.

177. G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.," *J. Comput. Chem.*, vol. 30, pp. 2785–2791, Dec. 2009.

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

179. R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

180. A. J. Li and R. Nussinov, "A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking.," *Proteins: Struct., Funct., Bioinf.*, vol. 32, pp. 111–127, July 1998.

181. A. H. Narten and H. A. Levy, "Liquid Water: Molecular Correlation Functions from X-Ray Diffraction," *J. Chem. Phys.*, vol. 55, no. 5, pp. 2263–2269, 1971.

182. D. Ringe and G. A. Petsko, "Study of protein dynamics by x-ray diffraction.," *Methods Enzymol.*, vol. 131, pp. 389–433, 1986.

183. P. A. Karplus and G. E. Schulz, "Prediction of chain flexibility in proteins," *Naturwissenschaften*, vol. 72, pp. 212–213, Apr. 1985.

184. D. E. Tronrud, "Knowledge-Based *B*-Factor Restraints for the Refinement of Proteins," *J. Appl. Crystallogr.*, vol. 29, pp. 100–104, Apr. 1996.

185. *Virtual Screening in Drug Discovery (Drug Discovery Series)*. CRC Press, 1 ed., Mar. 2005.

186. C. de Graaf, P. Pospisil, W. Pos, G. Folkers, and N. P. E. Vermeulen, "Binding Mode Prediction of Cytochrome P450 and Thymidine Kinase Protein-Ligand Complexes by Consideration of Water and Rescoring in Automated Docking," *J. Med. Chem.*, vol. 48, pp. 2308–2318, Apr. 2005.

187. C. de Graaf, C. Oostenbrink, P. H. J. Keizers, T. van der Wijst, A. Jongejan, and N. P. E. Vermeulen, "Catalytic Site Prediction and Virtual Screening of Cytochrome P450 2D6 Substrates by Consideration of Water and Rescoring in Automated Docking," *J. Med. Chem.*, vol. 49, pp. 2417–2430, Apr. 2006.

188. B. C. Roberts and R. L. Mancera, "Ligand-Protein Docking with Water Molecules," *J. Chem. Inf. Model.*, vol. 48, pp. 397–408, Jan. 2008.

189. R. Santos, J. Hritz, and C. Oostenbrink, "Role of Water in Molecular Docking Simulations of Cytochrome P450 2D6," *J. Chem. Inf. Model.*, vol. 50, pp. 146–154, Jan. 2010.

190. R. Thilagavathi and R. L. Mancera, "Ligand-Protein Cross-Docking with Water Molecules," *J. Chem. Inf. Model.*, vol. 50, pp. 415–421, Feb. 2010.

191. D. Bellocchi, A. Macchiarulo, G. Costantino, and R. Pellicciari, "Docking studies on PARP-1 inhibitors: insights into the role of a binding pocket water molecule," *Bioorgan. Med. Chem.*, vol. 13, pp. 1151–1157, Feb. 2005.

192. R. L. Mancera, "De novo ligand design with explicit water molecules: an application to bacterial neuraminidase.," *J. Comput.-Aided Mol. Des.*, vol. 16, pp. 479–499, July 2002.

193. M. Rarey, B. Kramer, and T. Lengauer, "The particle concept: placing discrete water molecules during protein-ligand docking predictions.," *Proteins: Struct., Funct., Bioinf.*, vol. 34, pp. 17–28, Jan. 1999.

194. M. L. Verdonk, G. Chessari, J. C. Cole, M. J. Hartshorn, C. W. Murray, Nissink, R. D. Taylor, and R. Taylor, "Modeling Water Molecules in Protein-Ligand Docking Using GOLD," *J. Med. Chem.*, vol. 48, pp. 6504–6515, Oct. 2005.

195. N. Huang and B. K. Shoichet, "Exploiting Ordered Waters in Molecular Docking," *J. Med. Chem.*, vol. 51, pp. 4862–4865, Aug. 2008.

196. G. Lemmon and J. Meiler, "Towards ligand docking including explicit interface water molecules," *PLoS ONE*, vol. 8, pp. e67536+, June 2013.

197. M. L. Raymer, P. C. Sanschagrin, W. F. Punch, S. Venkataraman, E. D. Goodman, and L. A. Kuhn, "Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm.," *J. Mol. Biol.*, vol. 265, pp. 445–464, Jan. 1997.

198. G. E. Kellogg, S. F. Semus, and D. J. Abraham, "HINT: a new method of empirical hydrophobic field calculation for CoMFA.," *J. Comput.-Aided Mol. Des.*, vol. 5, pp. 545–552, Dec. 1991.

199. D. L. Chen and G. E. Kellogg, "A computational tool to optimize ligand selectivity between two similar biomacromolecular targets," *J. Comput.-Aided Mol. Des.*, vol. 19, pp. 69–82, Feb. 2005.

200. A. Amadasi, F. Spyrakis, P. Cozzini, D. J. Abraham, G. E. Kellogg, and A. Mozzarelli, "Mapping the Energetics of WaterProtein and WaterLigand Interactions with the Natural HINT Forcefield: Predictive Tools for Characterizing the Roles of Water in Biomolecules," *J. Mol. Biol.*, vol. 358, pp. 289–309, Apr. 2006.

201. A. Amadasi, J. A. Surface, F. Spyrakis, P. Cozzini, A. Mozzarelli, and G. E. Kellogg, "Robust Classification of Relevant Water Molecules in Putative Protein Binding Sites," *J. Med. Chem.*, vol. 51, pp. 1063–1067, Feb. 2008.

202. A. T. García-Sosa, R. L. Mancera, and P. M. Dean, "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.*, vol. 9, pp. 172–182, June 2003.

203. M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, and C. W. Murray, "Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance," *J. Med. Chem.*, vol. 50, pp. 726–741, Feb. 2007.

204. H. Akaike, "A new look at the statistical model identification," *IEEE T. Automat. Contr.*, vol. 19, no. 6, pp. 726–723, 1974.

205. L. A. Kuhn, C. A. Swanson, M. E. Pique, J. A. Tainer, and E. D. Getzoff, "Atomic and residue hydrophilicity in the context of folded protein structures.," *Proteins: Struct., Funct., Bioinf.*, vol. 23, pp. 536–547, Dec. 1995.

206. J. Israelachvili and R. Pashley, "The hydrophobic interaction is long range, decaying exponentially with distance," *Nature*, vol. 300, pp. 341–342, Nov. 1982.

207. R. Wang, Y. Lu, and S. Wang, "Comparative Evaluation of 11 Scoring Functions for Molecular Docking," *J. Med. Chem.*, vol. 46, pp. 2287–2303, May 2003.

208. N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil, "Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go.," *Br. J. Pharmacol.*, vol. 153 Suppl 1, pp. S7–S26, Mar. 2008.

209. S.-Y. Huang, S. Z. Grinter, and X. Zou, "Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions," *Phys. Chem. Chem. Phys.*, vol. 12, no. 40, pp. 12899–12908, 2010.

210. H. G. Wallnoefer, K. R. Liedl, and T. Fox, "A challenging system: free energy prediction for factor xa.," *J. Comp. Chem.*, vol. 32, pp. 1743–1752, June 2011.

211. I. Maffucci and A. Contini, "Explicit Ligand Hydration Shells Improve the Correlation between MM-PB/GBSA Binding Energies and Experimental Activities," *J. Chem. Theory Comput.*, vol. 9, pp. 2706–2717, May 2013.

212. S. Krystek, T. Stouch, and J. Novotny, "Affinity and Specificity of Serine Endopeptidase-Protein Inhibitor Interactions," *J. Mol. Biol.*, vol. 234, pp. 661–679, Dec. 1993.

213. H. J. Böhm, "The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure.," *J. Comput.-Aided Mol. Des.*, vol. 8, pp. 243–256, June 1994.

214. K. L. Damm-Ganamet, R. D. Smith, J. B. Dunbar, J. A. Stuckey, and H. A. Carlson, "CSAR Benchmark Exercise 20112012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series," *J. Chem. Inf. Model.*, Apr. 2013.

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

216. J. W. Ponder and D. A. Case, "Force fields for protein simulations.," *Adv. Protein Chem.*, vol. 66, pp. 27–85, 2003.

217. R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell, "A semiempirical free energy force field with charge-based desolvation," *J. Comput. Chem.*, vol. 28, pp. 1145–1152, Apr. 2007.

218. S. Yin, L. Biedermannova, J. Vondrasek, and N. V. Dokholyan, "MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening," *J. Chem. Inf. Model.*, vol. 48, pp. 1656–1662, Aug. 2008.

219. G. Galliéro, C. Boned, A. Baylaucq, and F. Montel, "Molecular dynamics comparative study of Lennard-Jones -6 and exponential -6 potentials: application to real simple fluids (viscosity and pressure).," *Phys. Rev. E,*, vol. 73, June 2006.

220. A. Jain, "Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities," *J. Comput.-Aided Mol. Des.*, vol. 10, pp. 427–440, Oct. 1996.

221. M. R. McGann, H. R. Almond, A. Nicholls, J. A. Grant, and F. K. Brown, "Gaussian docking functions.," *Biopolymers*, vol. 68, pp. 76–90, Jan. 2003.

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

223. T.-C. Lim, "The relationship between lennard-jones (12-6) and morse potential functions.," *Z. Naturforsch.*, vol. 58a, pp. 615–617, 2003.

224. A. Bondi, "van der Waals Volumes and Radii," *J. Phys. Chem.*, vol. 68, pp. 441–451, Mar. 1964.

225. D. Seeliger and B. L. de Groot, "Atomic contacts in protein structures. A detailed analysis of atomic radii, packing, and overlaps.," *Proteins: Struct., Funct., Bioinf.*, vol. 68, pp. 595–601, Aug. 2007.

226. T. S. G. Olsson, M. A. Williams, W. R. Pitt, and J. E. Ladbury, "The Thermodynamics of ProteinLigand Interaction and Solvation: Insights for Ligand Design," *J. Mol. Biol.*, vol. 384, pp. 1002–1017, Dec. 2008.

227. G. Eugene Kellogg and D. J. Abraham, "Hydrophobicity: is LogPo/w more than the sum of its parts?," *European J. Med. Chem.*, vol. 35, pp. 651–661, Aug. 2000.

228. M. Eldridge, C. Murray, T. Auton, G. Paolini, and R. Mee *J. Comput.-Aided Mol. Des.*, no. 5, pp. 425–445.

229. M. W. Chang, C. Ayeni, S. Breuer, and B. E. Torbett, "Virtual Screening for HIV Protease Inhibitors: A Comparison of AutoDock 4 and Vina," *PLoS ONE*, vol. 5, pp. e11955+, Aug. 2010.

230. J.-C. Wang, J.-H. Lin, C.-M. Chen, A. L. Perryman, and A. J. Olson, "Robust Scoring Functions for ProteinLigand Interactions with Quantum Chemical Charge Models," *J. Chem. Inf. Model.*, vol. 51, pp. 2528–2537, Sept. 2011.

231. I. D. Kuntz, K. Chen, K. A. Sharp, and P. A. Kollman, "The maximal affinity of ligands," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 9997–10002, Aug. 1999.

232. N. Artemenko, "Distance Dependent Scoring Function for Describing Protein-Ligand Intermolecular Interactions," *J. Chem. Inf. Model.*, vol. 48, pp. 569–574, Mar. 2008.

233. T. Sato, T. Honma, and S. Yokoyama, "Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for in Silico Screening," *J. Chem. Inf. Model.*, vol. 50, pp. 170–185, Jan. 2010.

234. J. D. Durrant and J. A. McCammon, "NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes," *J. Chem. Inf. Model.*, vol. 50, pp. 1865–1871, Sept. 2010.

235. P. J. Ballester and J. B. O. Mitchell, "A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking," *Bioinformatics*, vol. 26, pp. 1169–1175, May 2010.

236. J. D. Durrant and J. A. McCammon, "NNScore 2.0: A Neural-Network Receptor-Ligand Scoring Function," *J. Chem. Inf. Model.*, Oct. 2011.

237. L. Li, B. Wang, and S. O. Meroueh, "Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries.," *J. Chem. Inf. Model.*, vol. 51, pp. 2132–2138, Sept. 2011.

238. R. Wang, X. Fang, Y. Lu, C.-Y. Y. Yang, and S. Wang, "The PDBbind Database: Methodologies and Updates," *J. Med. Chem.*, vol. 48, pp. 4111–4119, May 2005.

239. B. Kuhn, J. E. Fuchs, M. Reutlinger, M. Stahl, and N. R. Taylor, "Rationalizing tight ligand binding through cooperative interaction networks.," *J. Chem. Inf. Model.*, vol. 51, pp. 3180–3198, Dec. 2011.

240. H. Zou and T. Hastie, "Regularization and variable selection via the Elastic Net," in *J. R. Stat. Soc. B*, pp. 301–320, 2005.

241. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," in *Ann. Stat.*, vol. 29, pp. 1189–1232, 2000.

242. S. H. Sleigh, J. R. H. Tame, E. J. Dodson, and A. J. Wilkinson, "Peptide Binding in OppA, the Crystal Structures of the Periplasmic Oligopeptide Binding Protein in the Unliganded Form and in Complex with Lysyllysine," *Biochemistry*, vol. 36, pp. 9747–9758, Aug. 1997.

243. G.-B. Li, L.-L. Yang, W.-J. Wang, L.-L. Li, and S.-Y. Yang, "ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to ProteinLigand Interactions," *J. Chem. Inf. Model.*, Feb. 2013.

244. R. D. Head, M. L. Smythe, T. I. Oprea, C. L. Waller, S. M. Green, and G. R. Marshall, "VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands," *J. Am. Chem. Soc.*, vol. 118, pp. 3959–3969, Jan. 1996.

245. C. Kramer and P. Gedeck, "Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets," *J. Chem. Inf. Model.*, vol. 50, pp. 1961–1969, Nov. 2010.

246. D. M. Hamby, "A review of techniques for parameter sensitivity analysis of environmental models," *Environ. Monit. Assess.*, vol. 32, no. 2, pp. 135–154, 1994.

247. T. Cacoullos and V. Papathanasiou, "Lower variance bounds and a new proof of the central limit theorem," *J. Multivar. Anal.*, vol. 43, pp. 173–184, Nov. 1992.

248. S. F. F. Sousa, P. A. A. Fernandes, and M. J. a. J. Ramos, "Protein-ligand docking: current status and future challenges.," *Proteins: Struct., Funct., Bioinf.*, vol. 65, pp. 15–26, Oct. 2006.

249. C. S. Page and P. A. Bates, "Can MM-PBSA calculations predict the specificities of protein kinase inhibitors?," *J. Comput. Chem.*, vol. 27, pp. 1990–2007, Dec. 2006.

250. D. P. Oehme, R. T. Brownlee, and D. J. Wilson, "Effect of atomic charge, solvation, entropy, and ligand protonation state on MM-PB(GB)SA binding energies of HIV protease.," *J. Comput. Chem.*, vol. 33, pp. 2566–2580, Dec. 2012.

251. D. Zilian and C. A. Sotriffer, "SFCscoreRF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes," *J. Chem. Inf. Model.*, vol. 53, pp. 1923–1933, May 2013.

252. M. H. J. Seifert, "Targeted scoring functions for virtual screening," *Drug Discov. Today*, vol. 14, pp. 562–569, June 2009.

253. D. Fourches, E. Muratov, F. Ding, N. V. Dokholyan, and A. Tropsha, "Predicting Binding Affinity of CSAR Ligands Using Both Structure-Based and Ligand-Based Approaches," *J. Chem. Inf. Model.*, June 2013.

254. J. C. Faver, M. L. Benson, X. He, B. P. Roberts, B. Wang, M. S. Marshall, M. R. Kennedy, C. D. Sherrill, and K. M. Merz, "Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-Ligand Complexes," *J. Chem. Theory Comput.*, vol. 7, pp. 790–797, Feb. 2011.

255. K. M. Merz, "Limits of Free Energy Computation for Protein-Ligand Interactions," *J. Chem. Theory Comput.*, vol. 6, pp. 1769–1776, Apr. 2010.

256. J. C. Principe, *Information Theoretic Learning*. New York, NY: Springer New York, 2010.

257. G. Loomes and R. Sugden, "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty," *Econ. J.*, vol. 92, no. 368, pp. 805–824, 1982.

258. S. Verdú, "Mismatched Estimation and Relative Entropy," *IEEE T. Inform. Theory*, vol. 56, pp. 3712–3720, Aug. 2010.

259. T. Heskes, "Selecting Weighting Factors in Logarithmic Opinion Pools," in *Adv. Neur. In.*, pp. 266–272, 1998.

260. D. H. Wolpert, "The Lack of A Priori Distinctions Between Learning Algorithms," *Neural Comput.*, vol. 8, pp. 1341–1390, Oct. 1996.

261. D. H. Wolpert, "The supervised learning no-free-lunch Theorems," in *In Proc. 6th Online World Conference on Soft Computing in Industrial Applications*, pp. 25–42, 2001.

262. X. Zou, Yaxiong, and I. D. Kuntz, "Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model," *J. Am. Chem. Soc.*, vol. 121, pp. 8033–8043, Aug. 1999.

263. J.-H. Lin, A. L. Perryman, J. R. Schames, and J. A. McCammon, "Computational Drug Design Accommodating Receptor Flexibility: The Relaxed Complex Scheme," *J. Am. Chem. Soc.*, vol. 124, pp. 5632–5633, Apr. 2002.

264. J. L. Paulsen and A. C. Anderson, "Scoring Ensembles of Docked Protein:Ligand Interactions for Virtual Lead Optimization," *J. Chem. Inf. Model.*, vol. 49, pp. 2813–2819, Dec. 2009.

265. R. Amaro, R. Baron, and McCammon, "An improved relaxed complex scheme for receptor flexibility in computer-aided drug design," *J. Comput.-Aided Mol. Des.*, vol. 22, pp. 693–705, Sept. 2008.

266. F. Österberg, G. M. Morris, M. F. Sanner, A. J. Olson, and D. S. Goodsell, "Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock," *Proteins: Struct., Funct., Bioinf.*, vol. 46, pp. 34–40, Jan. 2002.

267. F. Ding and N. V. Dokholyan, "Incorporating Backbone Flexibility in MedusaDock Improves Ligand-Binding Pose Prediction in the CSAR2011 Docking Benchmark.," *J. Chem. Inf. Model.*, Dec. 2012.

268. J. Zhang, "The application of the Gibbs-Bogoliubov-Feynman inequality in mean field calculations for Markov random fields," *IEEE T. Image Process.*, vol. 5, pp. 1208–1214, July 1996.

269. R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, "Dissipation: The Phase-Space Perspective," *Phys. Rev. Lett.*, vol. 98, pp. 080602+, Feb. 2007.

270. X. Hu and W. H. Shelver, "Docking studies of matrix metalloproteinase inhibitors: zinc parameter optimization to improve the binding free energy prediction.," *J. Mol. Graph. Model.*, vol. 22, pp. 115–126, Nov. 2003.

271. X. Hu, S. Balaz, and W. H. Shelver, "A practical approach to docking of zinc metalloproteinase inhibitors.," *J. Mol. Graph. Model.*, vol. 22, pp. 293–307, Mar. 2004.

272. R. Schiffmann, A. Neugebauer, and C. D. Klein, "Metal-mediated inhibition of Escherichia coli methionine aminopeptidase: structure-activity relationships and development of a novel scoring function for metal-ligand interactions.," *J. Med. Chem.*, vol. 49, pp. 511–522, Jan. 2006.

273. T. Jain and B. Jayaram, "Computational protocol for predicting the binding affinities of zinc containing metalloprotein-ligand complexes.," *Proteins: Struct., Funct., Bioinf.*, vol. 67, pp. 1167–1178, June 2007.

274. U. F. Röhrig, A. Grosdidier, V. Zoete, and O. Michielin, "Docking to heme proteins," *J. Comput. Chem.*, vol. 30, pp. 2305–2315, Nov. 2009.

275. D. G. Myszka, Y. N. Abdiche, F. Arisaka, O. Byron, E. Eisenstein, P. Hensley, J. A. Thomson, C. R. Lombardo, F. Schwarz, W. Stafford, and M. L. Doyle, "The ABRF-MIRG'02 study: assembly state, thermodynamic, and kinetic analysis of an enzyme/inhibitor interaction.," *J. Biomol. Tech.*, vol. 14, pp. 247–269, Dec. 2003.

276. M. C. Jecklin, S. Schauer, C. E. Dumelin, and R. Zenobi, "Label-free determination of protein-ligand binding constants using mass spectrometry and validation using surface plasmon resonance and isothermal titration calorimetry.," *J. Mol. Recogn.*, vol. 22, no. 4, pp. 319–329, 2009.

277. J. Tellinghuisen and J. D. Chodera, "Systematic errors in isothermal titration calorimetry: Concentrations and baselines," *Anal. Biochem.*, vol. 414, pp. 297–299, July 2011.

278. C. Kramer, T. Kalliokoski, P. Gedeck, and A. Vulpetti, "The experimental uncertainty of heterogeneous public ki data," *J. Med. Chem.*, vol. 55, pp. 5165–5173, May 2012.

279. D. T. Jones, D. W. Buchan, D. Cozzetto, and M. Pontil, "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.," *Bioinformatics*, vol. 28, pp. 184–190, Jan. 2012.

280. T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, "Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing," *Cell*, vol. 149, pp. 1607–1621, June 2012.

281. T. Nugent and D. T. Jones, "Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis.," *Proc. Natl. Acad. Sci. USA*, vol. 109, pp. 9238–9239, June 2012.

282. X. Xia, E. G. Maliski, P. Gallant, and D. Rogers, "Classification of Kinase Inhibitors Using a Bayesian Model," *J. Med. Chem.*, vol. 47, pp. 4463–4470, Aug. 2004.

283. J. Besnard, G. F. Ruda, V. Setola, K. Abecassis, R. M. Rodriguiz, X.-P. Huang, S. Norval, M. F. Sassano, A. I. Shin, L. A. Webster, F. R. C. Simeons, L. Stojanovski, A. Prat, N. G. Seidah, D. B. Constam, G. R. Bickerton, K. D. Read, W. C. Wetsel, I. H. Gilbert, B. L. Roth, and A. L. Hopkins, "Automated design of ligands to polypharmacological profiles," *Nature*, vol. 492, pp. 215–220, Dec. 2012.

284. E. Fadda and R. J. Woods, "On the role of water models in quantifying the binding free energy of highly conserved water molecules in proteins: The case of concanavalin a," *J. Chem. Theory Comput.*, vol. 7, pp. 3391–3398, Aug. 2011.

285. M. A. Sahai and P. C. Biggin, "Quantifying water-mediated protein-ligand interactions in a glutamate receptor: a DFT study.," *J. Phys. Chem. B*, vol. 115, pp. 7085–7096, June 2011.

286. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "GROMACS: fast, flexible, and free.," *J. Comput. Chem.*, vol. 26, pp. 1701–1718, Dec. 2005.

287. B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations," *J. Comput. Chem.*, vol. 18, pp. 1463–1472, Sept. 1997.

288. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.

289. S. Nosé, "A molecular dynamics method for simulations in the canonical ensemble," *Mol. Phys.*, vol. 52, pp. 255–268, Nov. 1984.

290. W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Phys. Rev. A*, vol. 31, pp. 1695–1697, Mar. 1985.

291. M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J. Appl. Phys.*, vol. 52, pp. 7182–7190, Dec. 1981.

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