Computational Analyses of Protein-Ligand Interactions

Muhammad Kamran Haider

Submitted for the Degree of Doctor of Philosophy

University of York

Department of Chemistry

September 2010

Abstract

Protein-ligand interactions have a central role in all processes in living systems. A comprehensive understanding of protein interactions with small molecules is of great interest as it provides opportunities for understanding protein function and therapeutic intervention. The major aims of this thesis were to characterise protein-ligand interactions from databases of crystal structures and to apply molecular modelling techniques for accurate prediction of binding modes of molecular fragments in protein binding sites.

The first aspect of the project was the analysis of hydrogen bond donors and acceptors in 187 protein-ligand complexes of resolution 2.5Å or better. The results showed that an extremely small fraction of them were not explicitly hydrogen bonded, with the hydrogen bond criterion of donor-acceptor distance ≤ 3.5 Å and H-bond angle of ≥ 90°. It was also noticed that a vast majority of such cases were explicable on the basis of weak interactions and weak donor/acceptor strength. The results were consistent with reported observations for buried protein regions. In a series of docking calculations, the fraction of lost hydrogen bonds was evaluated as a discriminator of good versus bad docking poses. Docking and scoring with a standard program, rDock, did not create incorrect poses with missing hydrogen bonds to an extent that would make lost hydrogen bonds a strong discriminator. The second aspect of the research is related to weak (CH-π and XH-π, X=N,O,S) interactions. In a survey of IsoStar, a database of protein-ligand interactions, subtle differences were noticed in geometric parameters of π interactions involving different types of ligand aromatic rings with strong and weak donor groups in binding sites. The results supported the hypothesis that energetically favourable interaction patterns are more frequent when there are electron-donating substituents attached to the aromatic ring. Finally, the applicability of a modelling technique, multiple copy simultaneous search, in terms of predicting energetically favourable poses of solvents and fragments in target binding sites, was explored in detail. Several factors such as re-scoring with a better scoring function, use of multiple receptor structures and good quality prediction of water binding sites led to a robust protocol for high quality predictions of fragment binding in test datasets.

Acknowledgements

I am extremely grateful to my supervisor, Prof. Rod. Hubbard, for his guidance and encouragement throughout this project. I am very thankful for his support in academic matters in the university during last four years and for training towards long term goals, especially related to working in Pakistan at the end of PhD.

I would like to thank Dr. Hugues-Olivier Bertrand for his support, guidance and stimulating discussions during molecular modelling work.

I am particularly thankful to my friends in YSBL, Marcus Fischer, Javier Garcia, Yuan He, Justyna Korczynska and Rob Nicholls for providing me continuous support, encouragement and for enjoyable time in YSBL.

Special thanks to my very special friends Vikas Dangi, Chitvan Bochiwal and Hasan Basarir who made my stay in York a wonderful experience.

I would like to thank my parents for enabling me to think for myself and providing me an ideal environment for leaning.

I am hugely grateful to Higher Education Commission for providing financial support over last four years for my PhD studies.

Author's Declaration

Chapter 1, 2, 3 in this thesis is my own work.

Chapter 4 and 6 were done in collaboration with Dr. Hugues-Olivier Bertrand of Accelrys, who performed docking calculations with GOLD and validated my MCSS calculations by re-running them.

Chapter 5 in this thesis is my own work.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Protein-ligand interactions have a central role in all processes in living systems. A comprehensive understanding of protein interactions with small molecules is of great interest as it provides opportunities for understanding function and therapeutic intervention. Molecular recognition is, however, a complex interplay of several factors such as inter-molecular interactions of protein, ligand and the surrounding solvent, conformational variations of binding partners and the thermodynamics of molecular association. Over the past few decades experimental and computational techniques have been developed that shed light on the role of these factors. Our understanding of molecular recognition is still far from perfect. A brief literature review of some of the most important aspects of protein-ligand interactions is presented in this chapter followed by the motivation and primary aims of this research work.

1.1. Protein-ligand binding – Chemical and thermodynamic basis

The non-covalent reversible binding of small-molecules to proteins has a central role in biology. Several processes crucial to living systems involve specific recognition of small molecule ligands by proteins. For example, enzymes act on their substrates and catalyse key chemical reactions inside cells, transporters recognize specific molecules for their movement across membrane barriers, receptors specifically bind to hormones or other chemical messengers for inter- and intracellular communication and finally antibodies uniquely bind to foreign chemical agents to mount vital defence mechanisms against infection and disease. In general, the

binding of a protein with a ligand in an aqueous environment is given by the reaction:

$$P_{(aq)} + L_{(aq)} \leftrightharpoons PL_{(aq)}$$

The dissociation constant, K_D for this reaction is described as:

$$K_D = \frac{[P][L]}{[PL]}$$

Alternatively, the reciprocal of K_D or the association constant, K_A can be used. For a simple case of a ligand binding to a single site that is not affected by any other sites on the receptor, the value of K_D is the concentration of the ligand at which half of the binding sites are saturated[1]. K_D is therefore a measure of the affinity of the ligand towards its binding site and is measured in molar units, M.

Chemical reactions accompany a change in the free energy (ΔG) which is influenced by change in two other important quantities; enthalpy (ΔH) which is the heat content and entropy (ΔS) which is the temperature-independent degree of disorder. The resulting relationship between these quantities is given by:

$$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$$

The superscript '°' indicates the value of each of these properties at molar concentration of unity[2]. The change in free energy of binding is influenced by several factors such as electrostatic and van der Waals interactions, ionization effects, conformational changes and the role of solvent. All of these factors manifest themselves as favourable or unfavourable changes in entropy and enthalpy.

For example, the change in enthalpy is related to the breaking and formation of non-covalent interactions such as loss of protein-solvent and ligand-solvent hydrogen bonds and the formation of protein-ligand hydrophobic contacts and hydrogen

bonds. The relative strengths of these interactions determine whether or not enthalpy change is favourable[3].

In the same way changes in entropy upon binding are related predominantly to solvent displacement and reduction in conformational degrees of freedom. The burial of lipophilic surfaces results in an increase in entropy whereas confinement of the ligand and protein side-chains has an opposite effect. Furthermore, gain in enthalpy also accompanies an unfavourable change in entropy as formation of precise interactions causes structural rigidity and therefore decreases the entropy. This phenomenon is called enthalpy-entropy compensation[3, 4].

For a reaction to spontaneously occur, its free energy change should be negative. At equilibrium, ΔG° is related to the equilibrium constant by the following expression:

$$\Delta \mathrm{G}^{\mathrm{o}} = -\mathrm{RT}ln\mathrm{K}$$

where R is the gas constant and T is the absolute temperature. Using this relationship, free energy changes can be derived from experimentally measurable quantity, K_D. Biological K_D values exhibit a wide range from weak to very strong binding. Weak binding of coenzymes, such as nicotinamide, to enzymes is generally within, 0.1µM to 0.1mM. On the other hand, the strong binding of antigen-antibody complexes exhibits K_D values of up to 0.1 fM (1fM = 10^{-15}M)[1, 2]. In drug design, very low K_D values are desired because drugs can cause harmful side-effects due to off-target interactions. Therefore, binding affinity in the range of 0.1 to 10 nM is considered suitable[1]. The extremely high affinity is achieved by precisely engineering molecular interactions at the binding interface and improving the specificity of binding.

The specificity is conferred by inter-molecular interactions between the binding partners and their precise geometries. For example, electrostatic complimentarity is an essential feature of protein-ligand complexes. Similarly, hydrogen bond donors and acceptors are satisfied by their counterparts at the binding interface[2, 5].

1.2. Experimental methods for measuring binding affinity

The experimental techniques for measurement of binding affinity have been developed in past few years and are undergoing continuous development. The binding affinity can be determined indirectly using spectroscopic measurements such as change in absorption or fluorescence[6].

Indirect methods also involve physically restricting one of the binding partners and then measuring the free concentration of the other partner. Surface Plasmon Resonance (SPR) spectroscopy is a method based on similar principle and has been shown to reproduce binding affinity values that are consistent with more accurate direct measurements [7]. Other techniques for indirect measurement of protein-ligand binding affinity include NMR and Mass spectroscopy and atomic-force microscopy[6].

Direct methods give insight into thermodynamics of ligand binding and one of the most prominent technique is the isothermal titration calorimetry (ITC). ITC measures the heat of complex formation by incremental addition of ligand to the solution containing receptor molecules. The association constant, free energy change, enthalpic and entropic components can be derived from ITC data. Recent studies have highlighted the importance of thermodynamics analyses of ligand binding using ITC. For example, Klebe *et al*. showed that the minor difference in ΔG of a series of thrombin ligands was associated with significant mutually compensating changes in

enthalpic and entropic components of binding. In another study ITC measurements were used to identify the most optimal location of hydrogen bond donors and acceptors, for plasmepsin II inhibitors, by measuring enthalpic changes associated with different functional groups[8]. The importance of thermodynamic platforms based on ITC for studying protein-ligand binding has therefore been highlighted in recent reviews[4, 9].

1.3. Protein-Ligand Interactions

The non-covalent binding of small-molecule ligand to proteins is mediated by a variety of inter-atomic interactions. Mainly, these include electrostatic and van der Waals interactions (Figure 1.1). The affinity of receptor-ligand binding also heavily relies on contributions from other factors such as entropy, desolvation, flexibility of receptor structure and the structural water molecules in the binding site[6, 10]. In the following, a brief literature review of important protein-ligand interactions and other factors contributing to binding affinity is described.

1.3.1. Electrostatic Interactions

Electrostatic complimentarity between the protein and the ligand at the binding interface is vital for complex formation. The predominant types of electrostatic interactions include; hydrogen bonding, salt bridges, and metal interactions[6, 11] (Figure 1.1).

Hydrogen bonding is the most important directional interaction in biological macromolecules, known for conferring stability to protein structure and selectivity to protein-ligand interactions[12]. In general, hydrogen bonding occurs between two electronegative atoms, one of which (donor) has a covalently bound hydrogen atom

whereas the other (acceptor) has a lone pair of electrons. The strong electrostatic attraction arises from the attractive interaction between partial positive charge on the hydrogen atom and partial negative charge on the acceptor atom. Theoretical and experimental studies have confirmed an additional covalent component to hydrogen bonds as well which is based on the interaction between empty σ* anti-bonding orbital of the hydrogen atom and highest occupied orbital of the acceptor[12, 13].

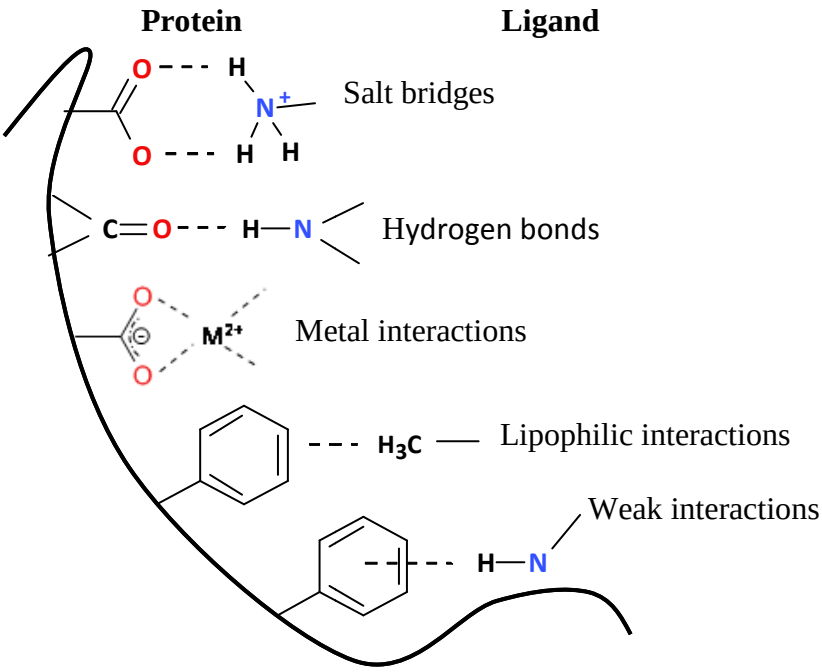


Figure 1.1. Major types of non-bonded interactions in protein-ligand complexes. (Adapted from Böhm, 2003[11])**.**

Experimental studies on hydrogen bonds based on X-ray scattering of ice crystals have indicated optimal donor-acceptor distance of 2.85Å and hydrogen-acceptor distance of 1.72Å[13]. Databases of non-bonded contacts in protein structures protein-ligand complexes, such as, IsoStar[14], SuperStar[15] and ReliBase[16] have been used to investigate geometric preferences of hydrogen bonding groups. In a recent survey, Bissantz *et al.* showed that the mean donor-acceptor distance of a typical hydrogen

bond (e.g., NH…CO hydrogen bond) lies in the range 2.8-3.1Å whereas the donor-hydrogen-acceptor angle was shown to be about 130°[10].

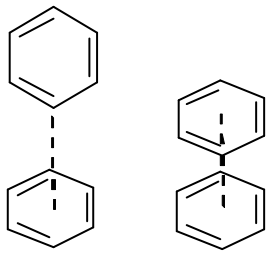
In the unbound state donors and acceptors from both protein binding site and the ligand are hydrogen bonded to water molecules. Upon binding the buried donors and acceptors at the interface make comparable hydrogen bonds. The difference in the strength of hydrogen bond in these two different environments (water and binding interface) determines the extent to which hydrogen bonds contribute to the affinity[6]. There have been a lot of studies to quantify the free energy gain from hydrogen bond formation but an agreement over a single value has not been reached. For example data from mutation studies on tyrosyl-tRNA synthetase suggest a reduction of binding affinity by 2.1-6.3 $kJmol^{-1}$ after removing a hydrogen bond between a strong donor and a strong acceptor[17]. Similarly, a buried hydrogen bond was shown to contribute about -5.4 $kJmol^{-1}$ to the stability of folded state of ribonuclease T1[18]. On the other hand free energy of desolvation of a hydrogen bond, as demonstrated by transfer of peptides from water to octanol, was show to be unfavourable (4.6 $kJmol^{-1}$). The differences observed in these values are compounded by the fact that the predominant electrostatic character of hydrogen bonds is influenced by the dielectric of surrounding medium. Buried hydrogen bonds in the protein interior are therefore considered to contribute more towards stability[6, 18]. In fact, a survey of buried NH and CO groups in protein structures indicated that only 1-2% of these group fail to form hydrogen bonds[19]. The satisfaction of hydrogen bonding potential within protein structures and at protein-ligand interface is further discussed in Chapter 2.

1.3.2. Hydrophobic Interactions

Hydrophobic interactions involve contacts between non-polar parts of the molecule (Figure 1.1). In protein-ligand complexes non-polar parts at the interacting surfaces are buried upon binding[10]. This causes displacement of water molecules thereby increasing the entropy. The hydrophobic interactions are therefore entropy-driven and have been shown to play crucial role in ligand binding[10, 11]. The relationship between burial of non-polar surface area and binding affinity is well established and amounts to an affinity gain of 30 cal mol^{-1} for 1 $Å^2$ of buried lipophilic surface area[11, 20, 21]. This implies that optimizing non-polar contacts of ligand atoms in hydrophobic protein pockets results in tighter binding. For example, Peters *et al.* demonstrated that optimizing interactions of aromatic rings in hydrophobic pocket of dipeptidyl peptidase IV resulted in 10^5-fold increase in affinity[22].

Aromatic residues in protein binding sites such as His, Phe, Trp and Tyr are frequently involved in aryl-aryl interactions[10]. Aromatic rings are known to interact with each other predominantly *via* one of two geometries: T-shaped edge to face and parallel displaced stacking interaction. Quantum mechanical studies on model systems such as benzene dimers have shown that these two geometries are isoenergetic[23]. In protein structures, however, parallel displaced geometry has been more frequently observed[24]. Aliphatic-aromatic interactions involving alkyl groups and aromatic rings are also commonly occurring interactions at non-polar interfaces[10]. Like aromatic-aromatic interactions, preferable interaction geometries include edge-to-face and parallel displaced interactions. The strength of aliphatic-aromatic interactions, particularly, CH-π interactions varies with increasing acidity of CH groups[25]. For example, *ab initio* calculation on model systems such as benzene

complexes with ethane, ethylene and acetylene show the highest interaction energy, -2.83 kcalmol^{-1}, for acetylene who has a more acidic CH group than others[26]. The role of weak interactions in protein-ligand complexes is discussed further in Chapter 3.



A B

Figure 1.2. Aryl-aryl interactions in protein structures and protein-ligand complexes. A. Edge-to-face geometry, B. Parallel stacking geometry.

1.4. Factors affecting protein-ligand binding affinity

In addition to electrostatic and shape complementarity between protein and ligands, there are some other important factors contributing to protein-ligand affinity. These are briefly discussed here.

1.4.1. Binding site water molecules

Water molecules play an important role in the structure and interactions of biomolecules. In the absence of a bound ligand, the binding site of a receptor is usually occupied by water molecules that are displaced upon ligand binding. Visualizing and characterizing water molecules in the binding sites on the basis of X-ray crystallographic structures is sometimes very difficult as these water molecules are highly disordered[10]. Highly conserved water molecules in the binding sites across multiple structures can however be considered to be tightly bound, for example water molecules in HSP90 N-terminal domain binding site are conserved across multiple structures (Figure 1.3). The displacement of water molecules increases the

entropy but it is offset by accompanying loss in enthalpy. The contribution of displacing a water molecule towards binding affinity therefore depends on how tightly it is bound and how efficiently the enthalpic loss by its displacement is compensated by interactions with the ligand molecule[6].

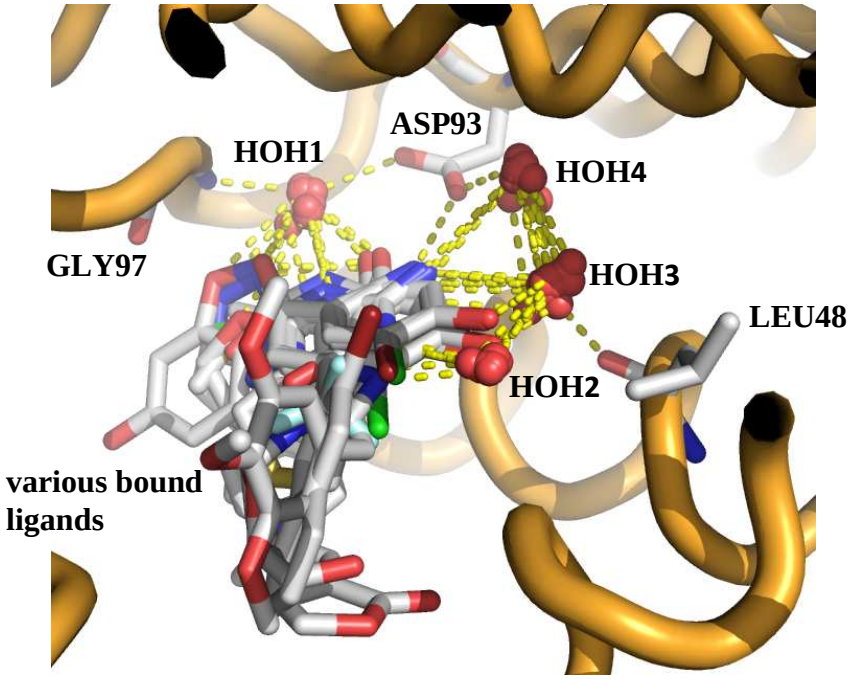


Figure 1.3. Water molecules at protein–ligand interface. The binding site of HSP90 ATPase domain is shown with ligands and water molecules superimposed from 11 different crystal structures (only one representative protein backbone is shown). Most ligands interact with tightly bound water molecules (HOH1, HOH3 and HOH4) which also interact with key binding site residues such as ASP93. In some cases, HOH2 is displaced upon ligand binding and replaced by a hydroxyl group.

Experimental studies to directly compare the effect of displacing a water molecule from the binding site indicate dependence on its 'coordination state' and the nature of compensating interactions[10]. For instance, it was shown that when water molecules formed two or less hydrogen bonds, its replacement with a close analog in the inhibitor molecule retains or improved the affinity of EGFR kinase and p38 MAP kinase inhibitors[27, 28]. On the other hand, it was noted for acetylcholinesterase

inhibitors that displacement of an extensively hydrogen-bonded water molecule resulted in significant loss in affinity[29].

1.4.2. Solvation and Desolvation

In addition to direct involvement in interactions with protein and ligand molecules, water also constitutes the medium in which biomolecular association occurs. Water molecules form a dynamic hydrogen bond network where each molecule is involved in 3 to 4 hydrogen bonds at a given instant[6]. The transfer of molecules that are non-polar or have a non-polar part causes disruption of this network and re-organization of water molecules around the non-polar solute molecules. The resulting unfavourable loss of entropy is compensated by stronger hydrogen bonds in the water molecules that are organized in clathrates. The complexation of non-polar molecules in water is therefore driven mainly by the release of the water molecules from the interface which increases entropy of the system. This has been termed as the classical hydrophobic effect and known as one of the driving forces for protein-ligand binding[6, 10, 11].

An almost opposite picture has emerged from spectroscopic studies such as neutron diffraction[30] and total internal reflection spectroscopy of aqueous solutions[31]. These studies indicate that water molecules at non-polar surfaces are not as rigidly bound and strong as considered by clathrate model[13]. In host-guest chemistry enthalpy-driven complexation has been observed giving rise to the idea of non-classical hydrophobic effect[23]. For example, the complex formation between benzene derivatives and a spherical hemicarcerand host was shown to be driven by favourable enthalpy change partially compensated by an unfavourable entropy change[32]. A classical example of this phenomenon from protein-ligand complexes

was reported by Homans *et al.*[33] where MUP (mouse major urinary protein) ligands were shown to bind with an extremely favourable enthalpy. The affinity of MUP ligands was shown to arise primarily from favourable solute-solute dispersion interactions with very little contribution from protein desolvation. This was attributed to sub-optimal hydration of the MUP binding site[10, 33].

The energetic cost related to desolvation plays an important role in drug design. An unfavourable desolvation of binding site or part of the binding site could result in full or partial loss of affinity[10]. For instance, Talhout *et al.* reported that the binding affinity of *p-tert*-butylbenzamidinium towards trypsin was lost because of unfavourable dehydration of Ser195 and His57 side-chains in oxyanion hole[34]. The alkyl substituent, which extended into the oxyanion hole, precluded the access of water molecules as shown by the calculated pKa shift for His57 side-chain. This resulted in a desolvation cost significant enough to lower binding affinity[34].

It is also well-established that the electrostatic interactions made by polar groups are shielded in an aqueous environment, which implies that the desolvation cost of a polar group in the ligand could be high enough to compensate for its potential interaction with a polar group in the binding site[6, 35]. This is also reflected in the fact that charge-assisted hydrogen bonds are not necessarily associated with increase in binding affinity[6].

1.4.3. Flexibility

The conformational flexibility of proteins is a well-known phenomenon and an important consideration in molecular recognition. Proteins are inherently flexible structures and conformational transitions of various scales play an important role in their function[36]. For example, the activity of HSP90 molecular chaperon is associated

with conformational transitions of an active site loop (residue 94-125 in yeast HSP90) which is also known as active-site lid[37] (Figure 1.3). Colombo *et al.* used molecular dynamics simulations to suggest that these conformational transitions vary with the nature of binding partner which could probably explain nucleotide-sensitive activity and modulation of HSP90[38].

Upon ligand binding, protein binding sites exhibit a variety of motions ranging from small-scale side-chain rearrangements to loop movements in the active site. In some cases, undefined protein structures undergo complete organization upon ligand binding[36]. Surveys of X-ray crystallographic structures of proteins in their apo and holo forms indicate backbone motions of up to 1Å in 20% of binding site residues[39] and 25% of binding sites[40].

X-ray crystallography, despite producing the highest number of experimentally solved protein structures, is limited in terms of studying conformational changes in proteins associated with the activity. A partial picture can, however, emerge from multiple structures representing functionally relevant conformational states. For example, conformational changes associated with active-site "lid" of the HSP90 N-terminal domain are suggested to be related to chaperone functioning of HSP90 based on the X-ray crystallographic structures of unliganded, ligand-bound and mutant structures of N-terminal domain[41]. These data suggest that multiple conformational states along HSP90 chaperone cycle can be targeted using structure-based drug design[42]. This is further supported by several structures of HSP90-ligand complexes emerging from drug design efforts that reveal altered conformational states of the active-site with different ligands[43, 44].

Nuclear magnetic resonance (NMR) is considered better suited to the study of structural dynamics of proteins. The main advantages are direct observation of protein in solution and the output in the form of an ensemble of low-energy conformations[36]. For example, data derived from ^{15}N - ^{1}H NMR experiments on mutant Sm14-M20, a *Schistosoma mansoni* fatty-acid binding protein, indicate differences in protein flexibility between apo and holo forms, particularly within the ligand binding region[36, 45].

Finally, computational approaches such as molecular dynamics (MD) simulations are used extensively to characterize protein flexibility[36]. Conformations generated from MD simulations can be used in other computational methods such as virtual screening, docking and scoring. Systems such as G-protein coupled receptors, (GPCRs) for which high resolution structural data is very challenging to generate using X-ray crystallography or NMR techniques, are widely studied using computational molecular dynamics. For instance, in one such study valuable predictive models were generated to successfully interpret QSAR data[46].

Continuing developments in these areas are expected to improve our understanding of the role of protein flexibility in protein function and molecular recognition[36].

1.5. Modelling and Prediction of Protein-Ligand Interactions

Despite the inherent complexity of molecular recognition, computational methods have been extensively developed over past few years for modelling and prediction of protein-ligand interactions. These approaches can be divided into three main categories[35]. The categories here are listed in decreasing order of accuracy and computational demand.

1.5.1. Methods based on Free Energy Calculations

Highly accurate modelling of protein-ligand binding is an extremely challenging task due to the complexity of the phenomenon. The principles of thermodynamics and statistical mechanics have been used to develop relatively accurate but computationally demanding treatment of protein-ligand interactions. These methods employ full-scale molecular dynamics simulation with explicit solvent and flexible protein and ligand molecules[6, 35]. Absolute or relative binding energies can be calculated in free energy calculation approaches. The absolute binding free energy calculations method, which are considered to be more accurate, involve separate simulation runs for solvated protein, ligand and the complex. No prior information on the structure and binding affinity of the complex is required[6]. In the case of relative free energy calculation methods, a known structure for the complex is used as reference and the difference in the binding free energy is calculated for the ligand of interest. The calculation is performed by alchemical transformation of reference ligand into target ligand. Molecular dynamics is used to exhaustively sample the configuration space[35]. The accuracy of these methods relies on the underlying atomic force field and the choice of an appropriate protocol for the problem at hand[11, 35].

1.5.2. MM-PBSA/GBSA Methods

MM-PBSA (Molecular Mechanics with Poisson-Boltzmann and Surface Area model) and MM-GBSA (Molecular Mechanics with Generalized Born and Surface Area model) methods were developed in 1990s and have been in practice and continuously developed since then[6, 35, 47]. These methods are based on the principle that free energy of binding can be decomposed into individual terms that describe

important contributions to binding[6]. The sum of energetic contributions from individual terms such as intra-molecular terms, van der Waals interactions, electrostatic interactions and solvation is calculated for protein, ligand and complex structures. These energetic terms are calculated from molecular mechanics force-fields such as CHARMm[48] and AMBER[49]. The energies are calculated as an ensemble average over conformations generated from MD simulation or simple energy minimization. Although conformations can be generated using explicit treatment of solvent, energy calculations are performed using implicit solvent models where solvent is represented as a continuum of high dielectric constant[35].

The polar interactions are therefore evaluated in the 'presence' of surrounding medium. Two major approaches in implicit solvent consideration are Poisson-Boltzmann (PB) equation and Generalized Born (GB) model. PB equation gives the most rigorous treatment but is computationally expensive. GB method is based on approximation to PB equation[50]. Further details on these two methods are described in Section 1.6.4. The entropic contribution to binding is calculated by normal mode analysis of MD trajectories however in most applications involving similar ligands, this is considered to be constant[35].

1.5.3. Docking and Scoring

Docking and scoring methods were designed to achieve high throughput computation of binding affinities. The accuracy is therefore less than the above mentioned methods. In general, docking and scoring involve generation of a set of poses for a ligand that can fit into a binding site. These poses are then rank ordered based on a scoring function[35]. A lot of scoring functions has been developed over the past few years. These can be classified depending on the approach that is used in

their formulation. Empirical scoring functions are based on the principle of additivity of individual terms contributing to total binding enthalpy[6]. These terms include important contributions to binding such as hydrogen bonds, hydrophobic interactions and ionic interactions and possibly entropic contributions. The weighted coefficients are adjusted to reproduce experimentally calculated binding affinities of a training dataset and derived using multiple linear regression and neural networks[11]. A generalized empirical scoring scheme can be described as:

$$\Delta G = \sum_i f_i \Delta G_i$$

where f_i and G_i are the coefficient and free energy associated with an interaction term *i*. Some of the pioneering examples include SCORE1[51] and Chemscore[52]. One of the disadvantages of these scoring functions is that due to additivity of terms, larger ligands get higher score than smaller ligands[53]. This could in some cases ignore the fact that large ligands could accompany higher entropic cost of binding because of more rotatable bonds. In some cases, this is circumvented by considering the number of rotatable bonds during entropic estimation in these scoring functions[54].

Another group of methods known as knowledge-based scoring functions are rooted in inverse Boltzmann law. The fundamental idea is that the frequency of the occurrence of a particular structural arrangement of two types of atoms is related to its energy. The database of protein structures can therefore be used to derive statistical potentials for a given atom pair which can then be converted to a potential of mean force. These scoring functions benefit largely from huge number of protein-ligand complexes to derive parameterization data but at the same time if a huge number of unique atom-atom pairs are included, there may not be statistically sound relationship for pseudo-potential of a particular atom-type[54]. Therefore, a careful

balance between well-defined atom-types and chemical diversity of interactions is treated. One such example is DrugScore[5] which uses 17 atom types based SYBYL mol2 format[54]. Other examples of knowledge-based scoring functions include BLEEP[55] and PMF [56].

The third class of scoring functions is known as force-field-based scoring functions. These methods use molecular-mechanics force-fields such as CHARMm[48] and Amber[49] to calculate enthalpy of binding. In most applications, values of non-bonded energy terms are pre-calculated on a grid and then interpolated to positions where atoms in docked protein-ligand complexes are located. The non-bonded terms normally include, van der Waals, electrostatic interactions and internal energy components related to bond lengths, bond angles, torsional angles. Additionally, a crude approximation of solvent effects by applying a distance-dependent dielectric constant can be added. More sophisticated approaches to account for long-range shielding of electrostatic interactions are implemented by combining Molecular Mechanics with Poisson-Boltzmann or Generalized Born approaches[54].

Force-fields contain parameters that allow for different atoms and their different properties. These parameterized are adjusted to reproduce experimentally or quantum-mechanically determined target data[57]. Therefore, quality of these parameters is extremely important in such calculations. Force-field based scoring could be time-consuming when applied on a large scale. Additionally, accurate modelling of charges is a considerable challenge[54].

Ferrara *et al.* showed in their assessment of most widely used scoring functions that the success rate of about 80% in terms of discriminating between correct poses among a set of decoys could be achieved for most of the scoring functions[58]. The

binding affinity prediction however remained challenging in most cases and at best a correlation coefficient of 0.51 was obtained for experimental and predicted values.

1.6. Protein-Ligand Interactions in Drug Discovery

The understanding of protein-ligand interactions and the ability to predict binding affinities are extremely important in drug discovery process. Despite their limitations, computational methods are extensively utilized in drug discovery campaigns. Structure-based drug discovery has become a vast field over the past few years because of continuous developments and several successful applications. The more recent trends, particularly connected to the contents of this thesis, are reviewed in the following. For instance, fragment-based lead discovery has emerged as one of the most successful approaches in drug discovery[59]. This has further increased the interest in computational methods for docking and scoring small molecules in protein binding sites.

1.6.1. Introduction to Fragment-based Lead discovery

Traditionally, drug discovery programs employ high-throughput screening (HTS) of huge corporate collections of compounds against the target of interest. Despite continuous developments, main challenges to HTS such as configuration of robust assays for screening number of compounds, insufficient coverage of chemical space and false-positive hits still remain[60]. Alternatively, fragment-based lead discovery (FBLD) methods for discovering potent lead compounds against drug targets have gained wide-spread interest in recent years. The size of fragments (molecular weight < 250 Da) allows a smaller library of compounds to sample a large chemical space and provide higher hit rates than screening of larger compounds as in high

throughput screening[60]. This is demonstrated by the fact that for fragments with up to 12 heavy atoms the chemical space is estimated to be 10^7 compounds whereas for drug-like compounds with 30 heavy atoms, this number increases up to 10^{60}. The small size and little complexity of fragments also enable them to bind with target sites more frequently with low affinity (100μM to 10mM range)[59]. Most importantly, the ligand efficiency for fragments, which is defined as the amount of free energy change per heavy atom upon binding, remains as good as for larger hit molecules. Consequently, it has been shown that optimizing 'hits' from fragment screening is a promising alternative[61,62].

Several recent reviews have summarised fragment-based lead discovery, and a number of examples have been published that demonstrate the various medicinal chemistry strategies to develop low-affinity fragments into high-affinity inhibitors against different targets[59, 63, 64, 65].

Due to their size and weak binding, conventional methods for detecting binding and activity are not useful in fragment screening. Alternatively, biophysical methods such as Nuclear Magnetic Resonance (NMR), Surface Plasmon Resonance (SPR) and X-ray crystallography are employed in order to detect and characterize fragment binding. The prioritization of fragments prior to experimental screening is particularly attractive[66]. Computational methods have been shown to assist FBLD at various stages including fragment-library design, docking and scoring of fragments for screening and lead-optimization[67].

1.6.2. Predicting functional group position in binding sites – GRID and MCSS

Algorithms to predict binding modes and affinity of small-molecule fragments and functional groups in binding sites were first developed before FBLD came into

practice. Most notably, two pioneering efforts in this case were the GRID program by Goodford[68] and Multiple Copy Simultaneous Search (MCSS) method from Karplus and co-workers[69]. Essentially, the idea behind these methods was to probe a protein active site for energetically favourable positions of polar, non-polar and charged functional groups of the ligand.

GRID represents chemical probes as single spheres. The interaction energy is calculated on a grid covering the target molecule or the binding site. The probes are mostly those molecules that can be represented as single spheres, such as water and methane, but 'multi-atom' probes can also be used by combing results from single-sphere probes[70]. The GRID energy function is an empirical scoring function which includes terms for non-bonded interactions, including van der Waals, electrostatic and hydrogen bonding interactions which are parameterized on the basis of equivalent terms in CHARMM energy function[68]. GRID has been successfully applied in the improvement of lead compounds and *de novo* ligand design[70, 71].

MCSS takes functional groups, fragments and even larger molecules (up to 30 atoms) as probes. Several copies of the probe (from 1,000 to 10,000) are randomly distributed in the binding site and energy minimization is performed simultaneously on all copies using time-dependent Hartree approximation[69]. This means that during minimization copies of the probe experience force-field only from the protein, independent of each other. The probe molecules are allowed to move under the effect of force-field and different copies converge to similar positions during minimization in which case duplicates are removed. At the end of minimization, a number of energy minima are obtained each of which is associated with an interaction energy and geometry. The CHARMm energy function is given in Table 5

(as molecular mechanics energy component, E_{MM}). The output from MCSS was used for the construction of ligands from functional group positions and binding modes using different strategies[72, 73] and applied on different protein targets such as HIV-1 aspartic proteinase[72], human α-thrombin[74], picornavirus capsid proteins[75] and others (see review by Schubert and Stultz[76]).

The X-ray crystallographic data on the binding of small probe molecules such as organic solvents to protein binding sites also became available based on a technique called Multiple Solvent Crystal Structures (MSCS)[77, 78]. The comparison of experimental positions of solvent probes with predicted positions from MCSS and/or GRID was performed for RNase A[79], thermolysin[80, 81] and elastase[82] which indicated good correlation in some cases but in others it highlighted the shortcomings of computational methods. Most of the poor quality predictions were associated with the lack of appropriate treatment of desolvation and conformational flexibility of the receptor[76].

In order to overcome some of the shortcomings, energy minima obtained from mapping algorithms can be evaluated based on rigorous scoring functions that estimate binding affinity based on implicit solvent models and therefore take the solvation component of binding free energy into account. In one such application, the binding free energy of MCSS minima was estimated based on contributions from bonding energy terms, van der Waals interaction and polar and non-polar solvation free energy[74]. The polar solvation free energy contribution was further divided into inter-molecular shielded electrostatic interactions, protein and ligand desolvation energy which was evaluated by solving linearized Poisson-Boltzmann (PB) equation. The ranking of energy minima produced in this was shown to be more realistic and in

agreement with experimental data. Solvation correction in this way is accompanied by additional computational cost[74, 76]. There have been significant developments in the use of implicit solvent methods in biomolecular simulations and estimation of binding affinity for protein ligand complexes[50]. Scoring schemes based on such methods have been employed in docking and virtual screening[47, 83-85].

1.6.3. Alternative computational approaches for solvent mapping and fragment docking

Other approaches have also been developed for computational solvent mapping. Dennis *et al.* developed CS-Map algorithm to address the problem of false-positives by using a scoring function that included a desolvation term[86]. The method was used to predict solvent positions and to identify consensus sites for seven enzymes[87]. A 'consensus site' can favourably interact with most of the probe molecules and is shown to be a major part of the binding site. The results from this study showed good agreement with experimental mapping results where such data were available. CS-Map was further developed into FT-Map which uses fast Fourier transform correlation approach to perform the initial search step[88]. The exhaustive evaluation of an energy function is done on rotational and translational grids for billions of conformations of probe molecules in a so called rigid body docking step. After the initial docking, the top 2000 poses for the probe molecule are further minimized using CHARMm potential with an analytic continuum electrostatics model. Finally, minimized poses are clustered ranked according to their Boltzmann averaged energies. The method was applied on elastase, for which solvent binding sites were identified in good agreement with experimental data. Similarly, for rennin, FT-Map was able to identify consensus sites that trace out the shape of inhibitor, aliskiren.

Multiple solvent crystal structure (MCSS) approach and FT-Map were used in conjunction on DJ-1 and Glucocerebrosidase and similar binding hotspots were identified[89].

Majeux *et al.* developed a method for docking and scoring small to medium-sized fragments in protein binding sites based on continuum electrostatic approach, hence named, Solvation Energy for Exhaustive Docking (SEED)[90]. Polar fragments are initially placed in the binding site making at least one hydrogen bond with optimal geometry. Apolar fragments are placed in hydrophobic region that are pre-evaluated for having low electrostatic desolvation and favourable van der Waals interactions with an uncharged probe sphere. The binding energy is then calculated for each fragment position where bad contacts are not present as a sum of van der Waals and electrostatic terms. The electrostatic term includes screened fragment-receptor interactions and fragment and receptor desolvation components which are calculated based on continuum electrostatic model[90].

In a later version of the SEED program, exhaustive electrostatic calculations are performed after a pre-processing step with approximated solvation treatment[91]. The so called electrostatic energy with fast solvation is based on linear distance-dependent dielectric model and Columbic approximation for electric displacement[92]. It was shown to speed up the docking and scoring process by discarding unfavourable binding modes and correct rank ordering for micro-molar inhibitors or close analogs of a set of proteins was also obtained. Further development of fragment docking methods in the Caflisch group includes flexible docking of combinations of three SEED-docked fragments using FFLD that uses a genetic algorithm and an efficient scoring function[93]. These triplets are then scored by a

variation of linear interaction energy model[94]. The application of these methods in development of low molecular inhibitors of four proteases and two kinases has been reviewed elsewhere[95].

Recently, another method for finding the most probable position and orientation of small ligands on protein surfaces has been developed[96] based on an integral equation theory of liquids, known as three-dimensional reference interaction site model (3D-RISM)[97, 98]. This method generates distribution functions of solvent site on a 3-D grid encompassing protein surface. The solvent is a mixture of water and target molecule for which spatial distributions of atomic sites are obtained. From the peaks of these distributions, favourable positions and orientations of ligand molecules are determined. The method was applied to thermolysin with solvent probes for which MSCS data are available[81] and sufficient agreement with experimental results was obtained[96].

1.6.4. Treatment of desolvation in solvent mapping/fragment docking

One of the most important issues in solvent mapping, fragment docking and protein-ligand docking in general is the accurate treatment of electrostatic interactions that take into account the solvent effects. The screened electrostatics interactions can be treated with different approaches or desolvation models. The methods developed after GRID and MCSS, as described earlier, address this issue in different ways. To date, continuum electrostatics models have been developed extensively and applied widely for protein-ligand binding affinity prediction. The relative success of such methods comes from the treatment of free energy of solvation using an implicit solvent model with continuum dielectric approach.

In a continuum dielectric implicit solvent model charged atoms in the solute molecules are embedded in a low-dielectric cavity, representative of the protein interior and are surrounded by a high dielectric continuum representing the characteristics of water[99].

The classical force-field based calculation of the total energy, E_T, of a molecule is based on the assumption that it can be decomposed into gas-phase potential energy, E_{gas}, and the free energy of solvation, ΔG_{solv},which is the free energy of transferring the molecule from vacuum to the solvent[100].

$E_T = E_{gas} + \Delta G_{solv}$

The implicit solvent model is then used to calculate the solvation free energy which is further divided into two sub-components.

$\Delta G_{solv} = \Delta G_{elec} + \Delta G_{np}$

ΔG_{elec} represents the electrostatic component which is the amount of free energy required to charge the molecule from zero to full charge in the presence of solvent. ΔG_{np} is non-polar component of salvation energy, which is the amount of free energy required to solvate the molecule when all charges have been removed.

The accurate estimation of the electrostatic potential used to calculate the energy is therefore essential to this approach. The most rigorous treatment is provided by Poisson theory which describes the electrostatic potential ϕ(r) as a function of the charge density ρ(r) and the position-dependent dielectric constant ε(r)[101]:

$$\nabla[\varepsilon(r)\nabla\phi(r)] = -4\pi\rho(r)$$

Additionally, charges from mobile ions can also be considered in the calculation of ϕ(r) from by assuming Boltzmann distribution of ions inside the potential field. The charge density ρ(r) in this case is equal to:

$$\rho(r) = \rho_f(r) + |e| \sum_j n_j z_j \exp[-\phi(r)|e|z_j] \Big/ kT$$

Where n_j and z_j are the molar concentration and charge of each ionic species j and |e| is the elementary charge[100].

Substituting the expression for ρ(r) into the Poisson equation gives the non-linear form of the Poisson-Boltzmann (PB) equation. A linearized form of PB equation is more commonly used in biomolecular modelling and simulation[102, 103].

$$\nabla[\varepsilon(r)\nabla\phi(r)] = -4\pi\rho(r) + \kappa^2\varepsilon(r)\phi(r)$$

The Debye-Huckel screening parameter, κ, encapsulates the screening effects of monovalent salt ions and is assumed to be 0.1Å^{-1} at physiological conditions. The electrostatic potential calculated from this equation is then used to calculate the ΔG_{elec} with the following expressions:

$$\Delta G_{elec} = \frac{1}{2}\sum_i q_i[\phi(r_i) - \phi(r_i)|_{vac}]$$

Due to the computational cost associated with this approach, alternative methods have been developed with efficient approximations. Generalized Born (GB) model is one such method which is based on Born expression for the solvation free energy of a single ion in a dielectric medium[50, 99]. A general expression for the electrostatic solvation energy under GB formalisms is given as[50, 99]:

$$\Delta G_{elec} = -\frac{1}{2}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right)\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{i,j} + \alpha_i\alpha_j \exp(-r_{i,j}^2 \big/ F\alpha_i\alpha_j)}}$$

where ε_p and ε_w represent solute and solvent dielectric constants, $r_{i,j}$ is the distance between atoms i and j, α_i is the GB radius of atom i. The factor F is a scaling factor for GB radii, whose most commonly used value is 4. The GB radius of an atom reflects

the distance of the atom from the solvent boundary[50, 99]. It was shown that in order to reproduce results obtained from PB equation, the calculation of Born radii or effective Born radii is crucial[104]. Poisson theory can be used to calculate the 'perfect' value for GB radii as it reproduces the results obtained from PB equation but the full advantage of GB formalisms lies in the alternative faster methods. Coulomb field approximation can be used to calculate GB radii for an atom with a charge at its centre[50, 105] and further developments have led to its application on biomolecules with off-centre charges[106, 107]. The definition of dielectric boundary is also of paramount importance in the accuracy of continuum dielectric models which in turn depends on the calculation of molecular surface. The exact calculation of solvent accessible surface represents a sufficiently accurate model for calculation of electrostatic potential with Poisson theory[100]. GB methods approximate molecular surface which leads to disagreement with the calculations based on PB equation. As a consequence much effort has been focused on approximating molecular surface that takes into account solvent excluded low dielectric cavities[50]. One such method, called Generalized Born using Molecular Volume (GBMV) uses a combination of Coulomb filed approximation and calculation of molecular surface to produce very good agreement with PB equation-based calculations[107].

The non-polar contribution to solvation free energy is computed based on the approximation that the van der Waals interactions between the solute and solvent molecules are proportional to the solvent accessible surface area of the solute [99]. An appropriate relationship for this is $G_{np} = \gamma SASA + b$, where γ and b are constants derived from experimental data and SASA is the solvent accessible surface area [50, 108].

1.7. Aims

This thesis is based on the research work undertaken under two broad themes; the analysis of protein-ligand interactions and prediction of energetically favourable positions of small-molecule fragments in protein binding sites.

The wealth of information in the databases of protein-ligand complexes can be used to carry out analyses that improve our understanding of molecular recognition. Similarly, docking calculations performing on a subset of high-resolution structures shed light on strength and weaknesses of scoring functions. The first part of the thesis (Chapter 2 and 3) is related to this aspect of the project, as explained in detail in the next section.

The second theme is covered in the second part (Chapter 4 and 5) of the thesis. With structure-based design gaining a very special role in drug discovery process, substantial efforts are devoted to the development and improvement of computational approaches for predicting protein-ligand interactions. Therefore, application of molecular modelling approaches to study the most challenging aspects of binding such as flexibility, water molecules, solvation and desolvation is an active and exciting area of research.

1.7.1. Unsatisfied Hydrogen Bonds

It has been noted that studying repulsive interactions in protein-ligand complexes is as important for our understanding of molecular recognition as the study of attractive interactions[11]. As noted previously, presence of unsatisfied hydrogen bond donors and acceptors in protein interior is a rare phenomenon. A buried donor or acceptor at protein-ligand interface should also cause destabilization[6]. The main

aims of this study were to conduct a survey of a set of protein-ligand complexes to study and characterize unsatisfied buried donors and acceptors in the binding interfaces. Additionally, the use of unsatisfied donors/acceptors as a metric to discriminate between good and bad docking poses was also investigated (Chapter 2).

1.7.2. Weak Hydrogen Bonds

Weak hydrogen bonds have been observed in small molecules and protein-ligand complexes and implicated in phenomena such as crystal packing, supramolecular assembly and molecular recognition. Non-bonded contact analysis and systematic surveys of structural databases are routinely used to study these interactions. In this study weak interactions between ligand aromatic ring acceptors and donors in protein-binding sites were analyzed to study the distributions of geometric features and the potential effect of ring substitutions on interactions geometries (Chapter 3).

1.7.3. Prediction of fragment positions in binding site

As noted in Section 1.4, the accurate prediction of functional groups in protein binding sites with favourable interactions and correct geometry is of fundamental importance in ligand design but represents a formidable challenge. Various approaches have been developed over the last three decades to address this problem. In this part of the project, the application of a CHARMm force-field based method, Multiple Copy Simultaneous Search (MCSS), on X-ray crystallographic structures of protein-fragment/solvent complexes is investigated. MCSS calculations were performed on solvent mapping dataset and compared with another solvent mapping technique (Chapter 4). Similarly, calculations were performed for fragment docking dataset but with additional rescoring steps to overcome the shortcomings of

MCSS. Additionally, various aspects of calculations such as comparison with other docking programs and the affect of multiple receptor structures on the success rate were also investigated (Chapter 5).

Chapter 2

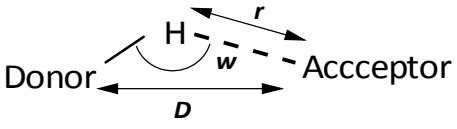
Unsatisfied Hydrogen Bond Donors and Acceptors at Buried Protein-Ligand Interfaces

2.1. Introduction

Hydrogen bonding is the most important directional interaction underpinning protein structure[109] and has been studied extensively using experimental and theoretical approaches. The role of hydrogen bonding in protein folding was first recognized in Pauling's proposals for secondary structure elements[110, 111]. The experimental determination of protein structures then provided the evidence which contributed to a better understanding of the role of hydrogen bonding in stability of protein structure and protein folding[19].

In one of the first benchmark studies of hydrogen bonding in proteins, Baker and Hubbard surveyed 12 globular proteins for hydrogen bonding patterns and characteristics[109]. In their analysis, they observed that approximately 11.2% and 12.4% of the main-chain CO and NH groups, respectively, were not explicitly hydrogen-bonded, and suggested that spatial constraints within the structure might prevent some of the carbonyl groups from satisfying their full hydrogen bonding potential. In a later study of hydrogen bonding in protein structures, McDonald and Thornton showed that these percentages might also include those groups that interact with disordered solvent and are expected to be satisfied[19]. They investigated the satisfaction of hydrogen bonding potential in a set of 57 high resolution protein structures (resolution 2.0 Å or better). They identified unsatisfied donors and

acceptors using different sets of criteria. In standard criteria, hydrogen-acceptor distance of 2.5 Å and hydrogen bond angle of 90°, 5.8% and 2.1% of main chain oxygen and nitrogen atoms, respectively, were observed to remain unsatisfied (Figure 2.1). In relaxed criteria, hydrogen-acceptor distance of 3.0 Å and hydrogen bond angle of 60°, these percentages reduced to 1.3% and 1.8%. They also showed that, in most of the cases, failure to form a hydrogen bond could be explained on the basis of factors such as electronic properties of the donor/acceptor type, the poor stereochemical quality of the structure, the difficulty in resolving some side-chains or compensating favourable interactions. Savage *et al.* surveys a set of 90 different protein structures at resolution 1.9 Å or better to study the correlation between the loss of potential hydrogen bonds and compensating stabilizing factors such as hydrophobic effect, ion pairs and disulfide bridges[112]. They noted that the loss of potential hydrogen bonds was highly correlated with buried surface area. Joh *et al.* observed in a set of six membrane protein structures that about 4% of polar atoms were not hydrogen bonded[113]. They also showed using double-mutant cycle analysis that hydrogen bonding interactions in membrane proteins were only modestly stabilizing.



Geometric criteria for hydrogen bonds:
$r \leq 2.5$Å
$D \leq 3.9$Å
$w \geq 90°$

Figure 2.1. Hydrogen bonding criteria from McDonald and Thornton[19]. *r*, hydrogen-acceptor distance, *D*, donor-acceptor distance, *w*, acceptor-hydrogen-donor angle.

In a more recent study, Fleming and Rose re-surveyed the unsatisfied groups in the McDonald study[19] and showed that almost all instances of unsatisfied donors or acceptor could either be explained or appeared because crystal structures in databases represented time-averaged snap-shots and therefore could fail static geometric tests for hydrogen bonding[114].

Hydrogen bonding also plays a major role in stabilizing protein-ligand complexes. It is, therefore, reasonable to expect that ligand binding should be accompanied by the satisfaction of the hydrogen bonding potential along the binding interface. The presence of unsatisfied buried hydrogen bonding groups in an apolar interface can reduce binding affinity significantly[10]. For example energetic penalty of up to 5 kcal mol^{-1} was reported for burial of a hydroxyl group in a hydrophobic pocket of mouse major urinary protein (MUP-I)[115]. The inclusion of such destabilizing terms has been highlighted as an important consideration in developing scoring functions for protein-ligand docking[116, 117]. For example, Reulecke *et al.* developed a scoring function HYDE that takes into account the unfavourable contribution resulting from dehydration of polar groups that fail to form hydrogen bonds with ideal geometry after protein-ligand binding[117]. The application of HYDE on test cases resulted in enrichment factors similar to or better than Flex[118] in 70% of the cases and worse in 30% of the cases. Similarly, in HINT forcefield, polar-polar interactions are either favoured or penalized depending on the charge and acid/base character of interacting atoms[119].

2.2. Aims

The aim of this study was to quantify the occurrence of unsatisfied or "lost" hydrogen bonds at protein ligand interfaces and assess whether such lost interactions could be used to discriminate between true and false ligand poses in computational docking. To achieve this, a set of 187 protein ligand complexes was analyzed to survey the number of unsatisfied hydrogen bond donors and acceptors in these binding sites. The same data set was then used in a docking study to assess whether unsatisfied hydrogen bonds could identify incorrect docking poses.

2.3. Methods

2.3.1. Dataset and Programs

The CCDC/Astex Test Set[120] was chosen as the data set to study the distribution of unsatisfied donors and acceptors in buried protein-ligand interfaces. This set contains 305 complexes chosen for evaluation of docking programs and other calculations related to protein-ligand interactions. It provides a diverse range of complexes at varying resolution with manually assigned protonation states for ionizable groups. An initial survey of the whole data set using protocols described below indicated a dependence of the number of unsatisfied hydrogen bonds on resolution. Therefore the final analysis covers a subset comprising 187 structures with resolution 2.5 Å or better.

The following protocols for various steps were written as Python scripts, using the MolKit package[121] for structure data parsing. For each complex in the data set, the ligand was identified and any atom belonging to the polypeptide, cofactors, metals

or water molecules lying within 7.0 Å of any ligand atom was considered to be part of the binding site.

2.3.2. Atom Typing

Following the extraction of ligands and binding sites, the next step was to type each polar atom according to its potential for being a hydrogen bond donor, acceptor or in some cases both.

Table 2.1. Typing of protein and ligand polar atoms with respect to their hydrogen bonding potential.

Proteins (atom names)		
Donor	**Acceptor**	**Donor-Acceptor**
Main-chain N	Main chain O	Ser OG
Asn ND2	Asp OD2, OD2	Thr OG1
Gln NE2	Glu OE1,OE2	Tyr OH
Trp NE1	Met SD	
Cys SG	Cys SG (in disulphide)	
Lys NZ	Asn OD1	
Arg NE, NH1, NH2	Gln OE1	
His ND1/NE2 (protonated)	His ND1/NE2	
Ligands (SYBYL atom types)		
Donor	**Acceptor**	**Donor-Acceptor**
N.am, N.4, N.pl3	N.1, N.ar1	N.3, N.2
	O.2, O.3, O.co2	O.3

Hydrogen atoms for protein polar atoms and protonation states had previously been assigned to the dataset[120], so atom typing was therefore straightforward for proteins. The atlas of side-chain and main-chain hydrogen bonding[122] was used to construct a simple dictionary of donor, acceptor and donor-acceptor groups (Table

2.1), with additional entries to reflect the protonation states of atoms in Asp, Glu and His residues. For Cys SG atoms, the presence of a disulphide bond was checked before typing.

SYBYL atom types for ligands were used to define donor/acceptor types as shown in Table 1. The atom types that can only have either donor or acceptor role were typed accordingly. The atom types that are able to switch roles between donor and acceptor (such as N.3) were assigned donor or acceptor type by checking for the presence of attached hydrogen atoms. Finally, the atom type O.3 (hydroxyl oxygen) was typed as both donor and acceptor.

2.3.3. Solvent Accessibility Calculations

Only donors/acceptors with zero solvent accessible surface area were chosen for subsequent identification of hydrogen bonds. Solvent exposed donors and acceptors that do not appear to be explicitly hydrogen bonded might still be interacting with disordered solvent molecules. Mobile solvent molecules are very difficult to visualize in X-ray structures therefore in order to prevent false-positives, only completely buried donor/acceptors atoms were investigated in this study. NACESS[123] was used to calculate accessible surface area (ASA) which is based on Lee and Richards algorithm[124]. ASA is calculated for each atom by rolling a probe of 1.4Å radius over the van der Waals surface of the complex (Figure 2.2). If the resulting ASA value is zero, only then it is considered as a completely buried donor or acceptor atom.

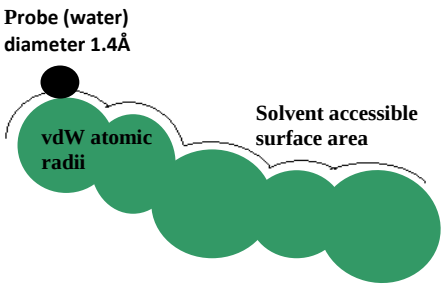


Figure 2.2 Solvent accessible surface calculation using Lee and Richards algorithm[124]. Surface traced by a solvent probe over van der Waals radii of atoms on target molecule is considered to be solvent accessible.

2.3.4. Optimization of side-chain orientations

Ligand binding sites in proteins could contain side-chains for which experimentally indistinguishable orientations are possible for example, Asn, Gln and His. This arises from the problem that due to the similar size of C, N and O atoms, they are very hard to distinguish in crystal structures. As Asn, Gln and His contain symmetrical side-chains therefore opposite orientations to those assigned in the structure could possibly occur. In the dataset Astex CCDC dataset, at least His side-chain orientations and their protonation states were manually assigned by authors[120].

For optimization of 'flip orientations' of Asn and Gln, the program REDUCE[125] was used. REDUCE assigns optimal flip states by optimizing hydrogen bonding and van der Waals overlapping. Additionally, it also optimized rotatable groups such as OH, SH, NH^{3+} and assigns an optimal orientation based on the same principle[125]. The updated coordinates of each PDB file are then used for further analysis.

2.3.5. Identification of hydrogen bonds

The hydrogen bonds formed by buried donors and acceptors from the protein and the ligand were identified using the relaxed criteria, defined by McDonald and Thornton[19], of donor-acceptor separation ≤ 3.5Å, hydrogen-acceptor distance ≤ 2.5 Å, hydrogen bond angle ≥ 60°. Occasionally, buried donor and acceptors make hydrogen bonds with interstitial water molecules at the protein-ligand interface. To allow for the disorder of such solvent molecules, the hydrogen bonding criteria is further relaxed by considering only the donor-acceptor distance at a cut-off value of 4.5 Å. A hydrogen bond is implied if a water molecule is found within the cut-off value, despite no explicit hydrogen bonding interaction.

2.3.6. Calculation of Normalized B factors

The normalized B factor values for individual atoms were derived according to the method of Parthasarathy and Murthy[126]. Donors and acceptors atoms were divided into hydrogen-bonded and unsatisfied groups. Within each group, for every atom, the B factor of corresponding Cα atoms was recorded. The mean B factor of Cα atoms within each group could this be calculated by:

$$b_\alpha = \sqrt{\frac{\sum b_i}{N}}$$

where b_i is B factor of *i*th Cα atom.

After obtaining b_α, the standard deviation, of donors and acceptors, about this mean value was calculated for each group as follows:

$$\sigma = \sqrt{\frac{\sum (b - b_\alpha)^2}{N}}$$

where *b* is the original B factor of the donor or acceptor atom and *N* is the total number of donors and acceptors in the group (hydrogen-bonded or unsatisfied).

The values of these two measures, b_α (mean B factor of C_α atoms) and σ (standard deviation about b_α) were used to calculate normalized B factor of each donor or acceptor, as follows:

$$B' = \frac{b - b_\alpha}{\sigma}$$

where *b* is the original B factor of the donor or acceptor atom.

The distributions of normalized B factors (B') across both groups, hydrogen-bonded and unsatisfied donors/acceptors, were then compared for analysis.

2.3.7. Ligand Docking

Ligands in the subset of 187 complexes were docked into their corresponding protein structures using rDock[127]. rDock uses a steady state genetic algorithm for docking search. The input ligand centre of mass is placed at the centre of binding site and then all rotatable bonds are randomized. The resulting poses are scored using an empirical scoring function. The scoring function is based on the terms that account for hydrogen bonds, attractive lipophilic interactions, repulsive steric interactions, positively-charged carbon-acceptor interactions, aromatic stacking interactions, donor-donor and acceptor-acceptor interactions and finally an estimate of entropy of ligand binding[127]. The overall score is a weighted sum of these terms and weighting coefficients are derived to reproduce experimental binding data.

In a variation from the standard scoring function, polar repulsive terms in the inter-molecular receptor-ligand interaction component are replaced by desolvation terms based on a weighted solvent accessible surface area model. Both the standard

scoring function and desolvation scoring function were used in separate docking runs.

Before docking, the conformation of each of the ligand structures was energy minimized using the *obminimize* utility of Open Babel[128] and then docked into the active site. The active site is defined as a predefined volume around crystal pose which is a spherical cavity with 10Å radius from the centre of ligand, excluding volume occupied by receptor atoms. For each ligand 20 docked poses were generated and the RMSD of each pose from the crystal pose was calculated. The fraction of unsatisfied ligand donors and acceptors was then calculated for each pose.

2.4. Results

2.4.1. Correlation with resolution

An initial survey of the complete data set of 305 protein-ligand complexes, suggested that the percentage of unsatisfied donors/acceptors increased as the resolution decreased. To characterize this, an average value for the percentages of unsatisfied main-chain NH and CO in binding sites was calculated for all the structures of a particular resolution. Figure 2.3 shows a plot of this average against the resolution range. The gradual upward trend can be attributed to the quality of structures affecting the accuracy of the calculations. Manual inspection of a few structures showed this was largely due to false positives where the unsatisfied donors/acceptors are just outside the cut-off criteria. Further analysis was, therefore, restricted to the 187 structures with resolution 2.5 Å or higher.

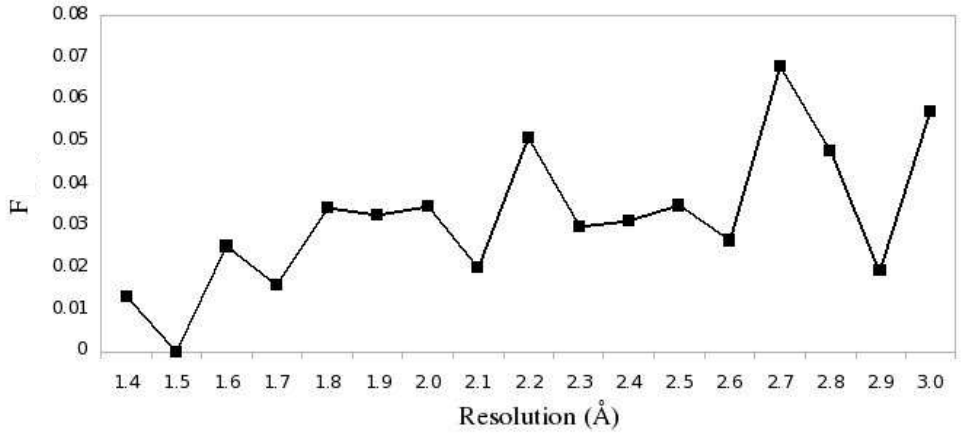


Figure 2.3. The average fraction of unsatisfied main-chain NH and CO groups, F, at different crystallographic resolutions. Crystal structures in each resolution value shown at X-axis were grouped together and their average fractions of unsatisfied main-chain NH and CO groups were plotted against the resolution.

2.4.2. The percentage of unsatisfied buried donors/acceptors

A total of 15, 542 polar groups at buried protein-ligand interfaces was surveyed in this subset of 187 structures. The percentage of unsatisfied groups for both proteins and ligands are shown in Table 2.2. Unsatisfied donors/acceptors appear roughly at similar frequencies in both proteins and ligands. The percentages of unsatisfied main-chain NH and CO groups (2.02% and 2.72%, respectively) are similar to those found in the McDonald and Thornton survey for protein core regions (1.3% and 1.8%, respectively). A higher percentage of unsatisfied protein donors/acceptors comes from side-chains atoms which has also been observed in the previous analyses of internal protein hydrogen bonding patterns[19, 114]. Table 2 shows the number of unsatisfied donors/acceptors across different types of side chain and the ligands. Overall, these results show that, as with the protein interior, losing a hydrogen bond in the binding site upon ligand binding is also an extremely low

incidence occurrence. For ligands, the major contribution of unsatisfied groups comes from polar atoms that are involved in weak interactions, most prominently, CH-O and NH-π hydrogen bonds.

Table 2.2. The percentage of buried unsatisfied donors/acceptors at protein-ligand interface.

Type	Total	Unsatisfied	%age
Main-chain donors	4453	90	2.0
Main-chain acceptors	4506	123	2.7
Side-chain donors	2469	67	2.7
Side-chain acceptors	2873	189	6.6
Ligand donors	428	5	1.9
Ligand acceptors	813	29	4.4

It was noticed that out of 15,542 total buried polar atoms surveyed, in 4,601 cases, either protein or ligand atoms were involved in hydrogen binding with interstitial the binding site also take part in polar interactions mediating the binding of ligand. One such example from chloramphenicol acetyltransferase complex with chloramphenicol is shown in Figure 2.4.

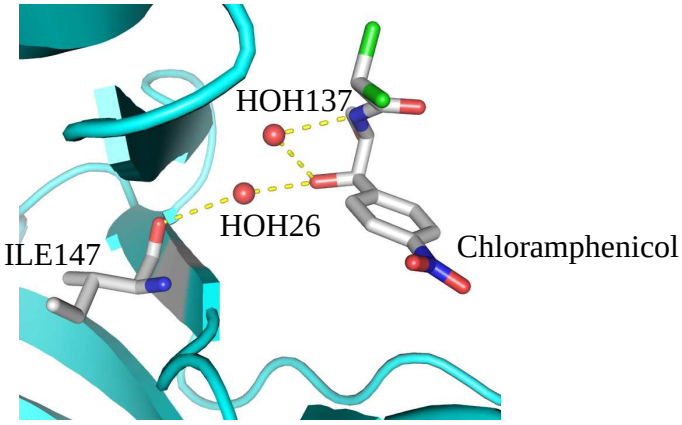


Figure 2.4. Water-mediated hydrogen bonds in chloramphenicol acetyltransferase-chloramphenicol complex (PDB code: 3CLA). Such water molecules in the binding site were considered during the identification of unsatisfied hydrogen bond donors and acceptors.

2.4.3. Normalized B factor profiles

B factors in protein structures reflect thermal fluctuation and positional disorder of atoms and therefore their analysis can provide information about protein stability and flexibility[129]. Normalized *B* (*B'*) factors can be calculated (as explained in Section 2.3.6) to compare among different categories of atoms in crystal structures. Previous analyses demonstrated that high-resolution structures show characteristic *B* factor profiles, with two peaks at -1.1 and 0.4, representing buried and exposed residues, respectively. The difference between the number of satisfied and unsatisfied donors and acceptors is very large, therefore, to compare the *B'* factor profiles, the number of satisfied donors/acceptors was scaled down to reflect similar sample size. The scaling factor is simply the ratio of satisfied to unsatisfied atoms. The resulting plots of *B'* factor distribution are shown in Figure 2.5. The similar *B'* factor profiles suggest that the occurrence of unsatisfied donors/acceptor surveyed in this study is not associated with disorder or flexibility of the structures.

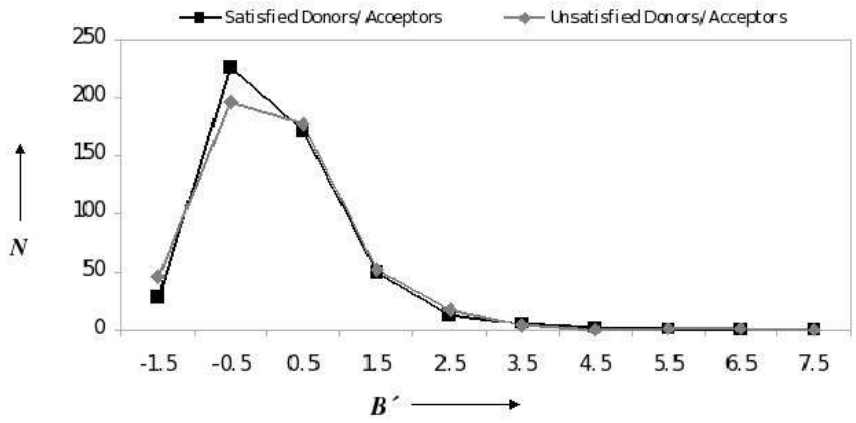


Figure 2.5 Normalized *B* factors, *B´*, for satisfied and unsatisfied donors and acceptors. For each group of atoms *B'* factors were calculated for comparison, as described in 'Methods'.

2.4.4. Docking Results

The success rate of docking programs is usually reported as the percentage of test cases in the dataset for which the top scoring pose has an RMSD ≤ 2.0 (or 3.0Å) from the crystallographically determined binding pose[58, 120]. The dataset used in this study is an extension of the original GOLD validation dataset. The success rate of GOLD on the original validation set was reported to be 71% at RMSD ≤ 3.0Å[120]. In this study, using rDock standard and desolvation scoring functions, we observe success rates of 65% and 66%, respectively. The scoring function has little effect on docking output; therefore all subsequent analysis is for poses generated using the desolvation scoring function.

The native binding pose is amongst the 20 poses generated by rDock for more than 90% of the dataset. However, these do not all receive the highest score, highlighting that improved scoring could increase docking performance. Figure 2.6 shows a plot of the fraction of unsatisfied donors and acceptors against the RMSD between the top scoring docked poses and crystallographically observed poses, presented for proteins, ligands and both proteins and ligands. The large number of points on the horizontal axis above RMSD of 3Å, shows that a substantial number of the top scoring poses have no unsatisfied hydrogen bonds (F = 0.0), but are docked far from the crystallographically observed position. This is emphasized in Table 3. This summarizes the number of complexes for which the poses have different fractions of lost hydrogen bonds for different RMSD values. The number of X-ray poses with different fractions of lost hydrogen bonds is also listed.

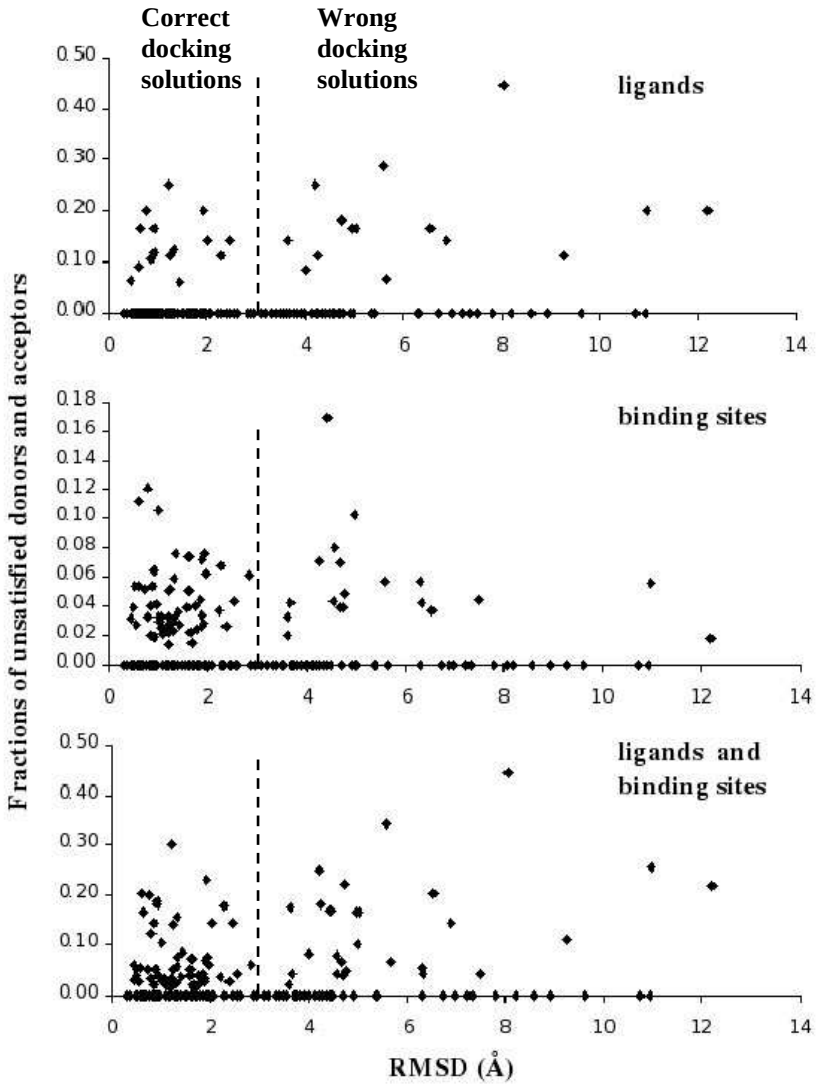


Figure 2.6. Fraction of unsatisfied donors and acceptors observed in top-scoring docking poses. Fractions are plotted (for ligands, protein binding sites and for both) against RMSD of the top scoring pose from the crystallographically determined pose. The scatter plots indicate that the fractions are somewhat similar for correct and incorrect docking solutions (marked by 3.0Å RMSD cut-off from the X-ray pose).

2.5. Discussion

2.5.1. Energetics of Lost hydrogen bonds

Intra-peptide hydrogen bonds are important for the stabilization of protein structures but there has been some debate about the energetic contribution made by each hydrogen bond[114]. For example, calorimetry experiments by Sholtz *et al.*

indicate that the enthalpy of alanine helix formation in water was estimated about 1 kcal mol^{-1} for a hydrogen bond[130], whereas the enthalpy of an intrapeptide hydrogen bond was estimated about 12 kcal mol^{-1} by Makhatadze and Privalov[131]. Myers and Pace estimated a net conformational stability of -1 to -2 kcal mol^{-1} per intrapeptide hydrogen bond for proteins by making series of single residue polar to non-polar mutations[132]. Similarly, buried hydrogen bonds have been estimated to contribute as much as -3.5 kcal mol^{-1} [18]. Despite the variation in these estimates of net energetic contribution, it is clear that intrapeptide hydrogen bonds stabilize protein structure and their loss will therefore result in conformational instability[114].

The results of our analysis of lost protein hydrogen bonds at the interface of protein-ligand interactions correlate closely with those of McDonald and Thornton for internal protein hydrogen bonds. In both cases, between 1-3% of main chain NH or CO hydrogen bonds are not made. We can apply the Boltzmann hypothesis to relate the probability of the occurrence of structural interactions to free energy[133], using the following expression:

$$p = e^{\Delta E_{hb} / RT}$$

Considering various estimates for the energetic cost of losing a buried hydrogen bond, this equation gives a probability ranging from 0.00273 (ΔE_{hb} = -3.5 kcal mol^{-1}) to 0.038 (ΔE_{hb} = -2 kcal mol^{-1}). In other words, the percentage of unsatisfied groups is expected to fall in the range of 0.27% to 3.8%, comparable to the total percentage of unsatisfied donors/acceptors observed in this study (3.23%, Table 2.2). If we consider only main-chain NH and CO groups to be true representatives of the population of unsatisfied donors/acceptors, the percentage falls to 1.37%. The low incidence of buried unsatisfied hydrogen bonding groups in both protein core

regions and buried protein-ligand interfaces is therefore consistent with each hydrogen bond contributing a few kcal mol^{-1} to the stability of the system.

2.5.2. Identification of lost hydrogen bonds

The positioning of H-atoms is crucial in identifying unsatisfied donors and acceptors. The identification of unsatisfied hydrogen bond donors and acceptors is complicated by factors such as ambiguity in the position of hydrogen atoms, experimentally indistinguishable side-chain orientations and the presence of disordered solvent. For groups with unambiguous hydrogen positions such as backbone NH (Fig 4A, PHE442 main-chain NH PDB code: 1tka) the standard criterion for identifying hydrogen bonds is most sufficient to identify hydrogen bonds. In such cases, the position of the hydrogen atom can be inferred from the peptide bond geometry. The position of the hydrogen atom can, however, be ambiguous (Fig 2.7B, TYR448 OH PDB code: 1tka) and it becomes quite difficult to ascertain hydrogen bonding partners. For example, in Figure 2.7B, the optimization of the side-chain orientation of Tyr448 favours hydrogen bonding to a nearby carbonyl oxygen.

For side-chains where alternative 'flip' orientations are possible, it is necessary to evaluate them thoroughly. An alternative orientation may satisfy hydrogen bonding but could create other problems such as steric clashes. Such cases were largely eliminated after the optimization of hydrogen atom positions and the orientation of flippable side-chains with REDUCE[125]. As REDUCE optimizes hydrogen bonding network therefore the orientations with a better hydrogen bond geometry are favoured.

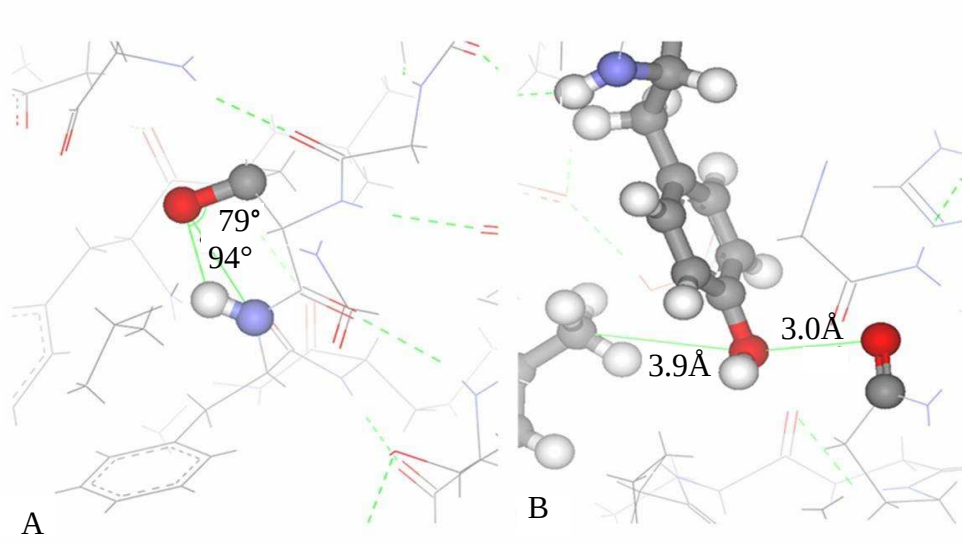


Figure 2.7. Identifying unsatisfied hydrogen bond donors and acceptors. A. A simple case, where H-atom position can be considered reliable, simple geometric criteria are useful to identify an H-bond (e.g., placement of H-atom in H-bond angle of 94° (which is within the cut-off used in this study). B. However in a case of ambiguous H-atom position, initial placement of hydrogen doesn't completely fulfil H-bond criteria (3.0Å distance but unsuitable angle). These cases were dealt with the H-bond optimization step in the protocol (See 2.3.4).

2.5.3. Types of unsatisfied donors and acceptors

The distribution of different protein and ligand polar atoms in the population of unsatisfied donors and acceptors is shown in Figure 2.8. The number of each of the different donors/acceptors found unsatisfied is normalized against the frequency of their occurrence at protein-ligand interfaces as $N_i = n_i \times f_i$ where n_i is the number of donor/acceptor i appearing unsatisfied as a fraction of the total unsatisfied donors/acceptors and f_i is the frequency of occurrence of donor/acceptor i in the binding site.

Previous surveys have shown that the presence of unsatisfied donor or acceptor in protein core regions can mostly be rationalized on the basis of either limitations in experimental/theoretical methods or due to compensating weak interactions[19, 109, 112, 114]. In this survey for buried protein-ligand interfaces, we observe that the largest category of unsatisfied donors/acceptors is side-chain atoms (54%) (Figure 2.8). Unsatisfied acceptors appear more frequently than unsatisfied donors. It has been observed that an unsatisfied donor at protein-ligand interface is energetically more unfavorable than an unsatisfied acceptor[19]. It was observed for a set of kinase inhibitors that one of the NH groups located at the centre of the hinge β-strand was almost always hydrogen bonded to the ligand with the exception of one case. On the other hand, ligands binding to hinge region of kinases often failed to hydrogen bond with a backbone carbonyl group leaving it unsatisfied[10].

This trend was also observed throughout the data set including protein backbone, side-chain and ligand atoms. Within the unsatisfied side-chain acceptors, Met SD and Tyr OH make the largest contribution (48%) (Figure 2.8). This can be explained on the basis of the lower electronegativity of sulfur atoms and delocalization of Tyr OH electrons over aromatic side-chains, making these two groups poor hydrogen bond acceptors. Similarly, ambiguity in the position of His, Asn, Gln side chains, steric crowding of Ser and Thr side-chains have been given as reasons for their inability to satisfy hydrogen bonding potential[19, 114]. We observe that the frequency of unsatisfied atoms belonging to these side-chains is consistent with these arguments.

Fleming and Rose compared electron density maps with the known PDB structures (which were documented by McDonald and Thornton as containing unsatisfied hydrogen bonds) and observed that in some cases the unsatisfied group has an

alternative but satisfied rotamer or a neighbouring side-chain's alternative rotamer allowing solvent access, particularly if it lies in a region of low electron density[114]. Such detailed analysis could further explain the frequency of appearance of some of the unsatisfied donors/acceptors. However, given the number of structures surveyed in this study, more detailed analysis of the unsatisfied donors/acceptors was not feasible.

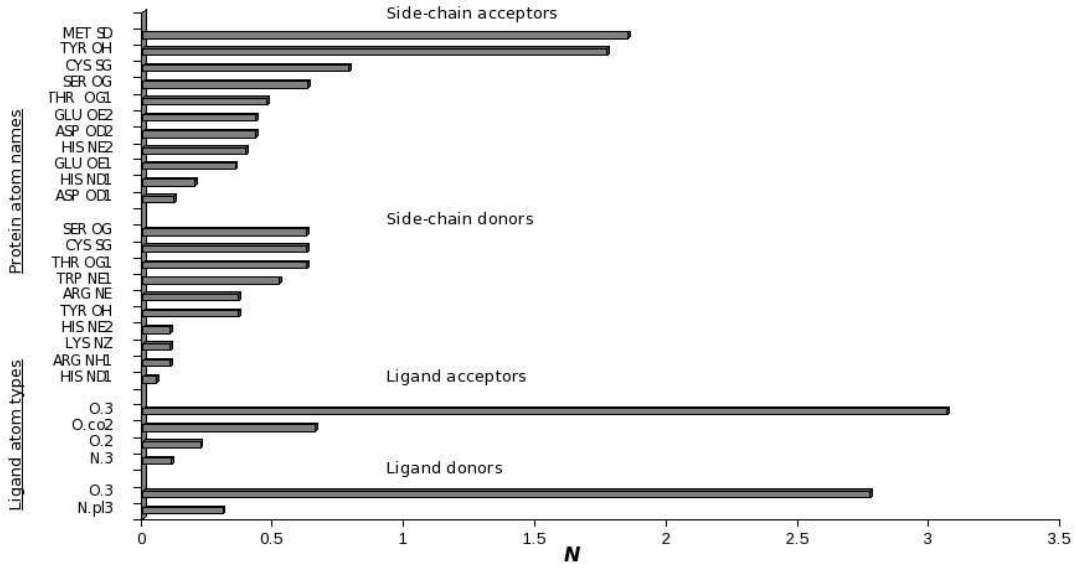


Figure 2.8. Distribution of unsatisfied donors/acceptors atom in side-chains and ligands. N is the number of unsatisfied occurrences for a given donor/acceptor normalized by the frequency of its occurrence at protein-ligand interfaces.

The unsatisfied donors and acceptors in ligands (Table 2.4) were also often associated with compensating weak interactions (Figure 2.9). Among ligand atoms, O.3 atom type (hydroxyl or ether oxygen) was observed to be most frequently unsatisfied and was mostly involved in CH-O interactions. An example of such weak interaction is shown in Figure 2.9, PDB code: 1poc. In this case, two oxygen atoms in the ligand were observed to be in hydrogen bonding geometries with CH groups in proteins. It has been noticed in previous studies that alkyl and aromatic groups

interact with groups containing O atoms in a similar manner as donor acceptor groups in conventional hydrogen bonds[134]. CH-O hydrogen bonds have also been implicated in protein-ligand complexes of pharmaceutical importance such as retinoic acid receptor complex with a selective agonist SR11254[135].

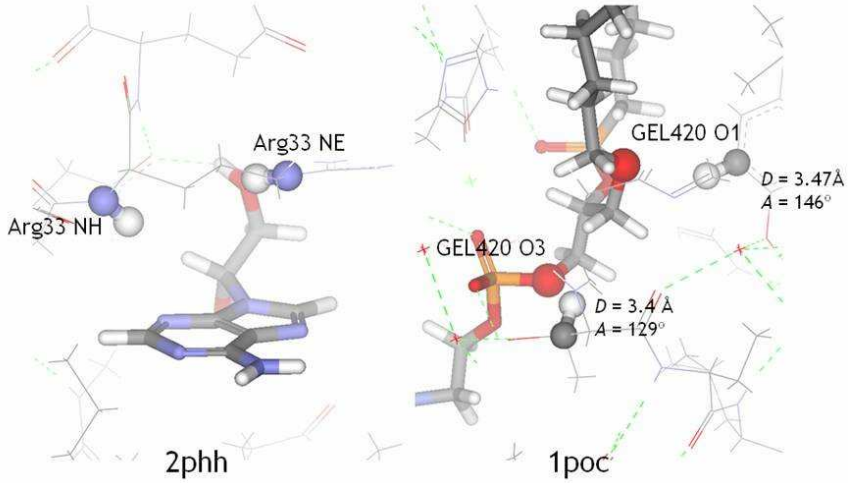


Figure 2.9. Weak interactions in unsatisfied donors and acceptors. Unsatisfied donors/acceptors often make weak interactions such as CH-O hydrogen bonds with geometric characteristics similar to conventional hydrogen bonds, as indicated by the values of D (donor-acceptor distance) and A (hydrogen bond angle) in the case of 1poc.

A further visual inspection of all these cases indicated that some of these cases could be rationalized on the basis of solvent disorder or orientation of rotatable groups. The criterion for donor-acceptor separation was extended by 1.0Å where a water molecule was present in the vicinity of a donor or acceptor, to take into account residual mobility associated with water molecules.

2.5.4. Unsatisfied buried donors/acceptors in protein-ligand docking

The extremely low incidence of unsatisfied buried donors and acceptors in X-ray structures prompts the argument that a correctly docked ligand pose should satisfy almost all of its hydrogen bonding potential at the binding interface. Since

imperfections in scoring functions lead to poor ranking of candidate poses, we asked the question whether poor poses generated by computational docking methods could be identified on the basis of the number of lost hydrogen bonds. A set of docking poses was generated for each of the set of 187 protein-ligand complexes. The results were somewhat surprising. The success of docking is measured as the RMSD between the crystallographically observed pose and that generated by the docking. The plots in Figure 2.6 show there was no substantial difference in the distribution of protein or ligand hydrogen bonding groups that were left unsatisfied. What is striking, however, is how many of the poses that have poor RMSD (>3Å) have no unsatisfied hydrogen bonds - the points that lie on the horizontal axis. This is emphasized by the summary provided in Table 2.3. The number of complexes for which docked poses have unsatisfied hydrogen bonds is approximately equivalent to the pattern of unsatisfied hydrogen bonds seen in the X-ray structures. For example, 18 complexes have an X-ray pose with between 10 to 25% of the buried protein-ligand hydrogen bonds unsatisfied, compared to 16 in the docked poses. In addition, very few of the docked poses with RMSD > 3Å have unsatisfied hydrogen bonds.

Table 2.3. A summary of the number of complexes with different fractions (F) of unsatisfied ligand donors and acceptors in X-ray binding modes and in docked ligand poses (subdivided into RMSD categories)

		F = 0	F < 0.1	0.1 < F < 0.25	F > 0.25
X-ray poses		151	10	18	8
Docked poses RMSD (Å)	**0.0 – 2.0**	83	7	11	7
	2.0 – 3.0	23	1	2	1
	>= 3.0	42	5	3	2

Three further analyses are useful to report. First, it is reassuring that the 83 docked poses for which the fraction of unsatisfied acceptors or donors (F) is 0 and where the RMSD (R) is less than 2.0 Å are all complexes where F = 0 for the X-ray pose. Secondly, 25 of the 39 complexes where the docked poses have F > 0 also have F > 0 in the X-ray pose. Of these 25 complexes, 16 have R < 2Å, the remaining 9 have R > 2Å. Finally, among the 14 cases where R > 2Å and F > 0, there were only 3 cases where the docked pose had a higher F than the X-ray pose.

An example from this last set (R > 2Å and F > 0) illustrates the issue with using unsatisfied hydrogen bonds as a criteria to select the correct pose. Figure 2.10 is a diagram of protein-ligand interactions in the complex 1poc. In the crystal structure, there are two oxygen atoms in the ligand (O1 and O3), that are buried and do not appear to form hydrogen bonds. For at least one of these O atoms there are CH groups in a position where weak hydrogen bonds could be formed. In docking, the top-scoring pose has an RMSD of 7.3 Å with a higher fraction of unsatisfied ligand donors and acceptors than in the X-ray structure. In addition to O1 and O3, three other buried acceptor atoms in the ligand, O2, O1P and O5P do not make hydrogen bonds.. The candidate pose from docking with the lowest RMSD from X-ray pose is ranked 12th. In this pose, one of the additional three unsatisfied acceptors (O5P) is hydrogen-bonded with one of the donor groups within the ligand. However, there are other poses where there are only three unsatisfied hydrogen bonds (ranked 7th and 19th in the docking), but which have a poor RMSD (4.8 Å and 9.5 Å, respectively).

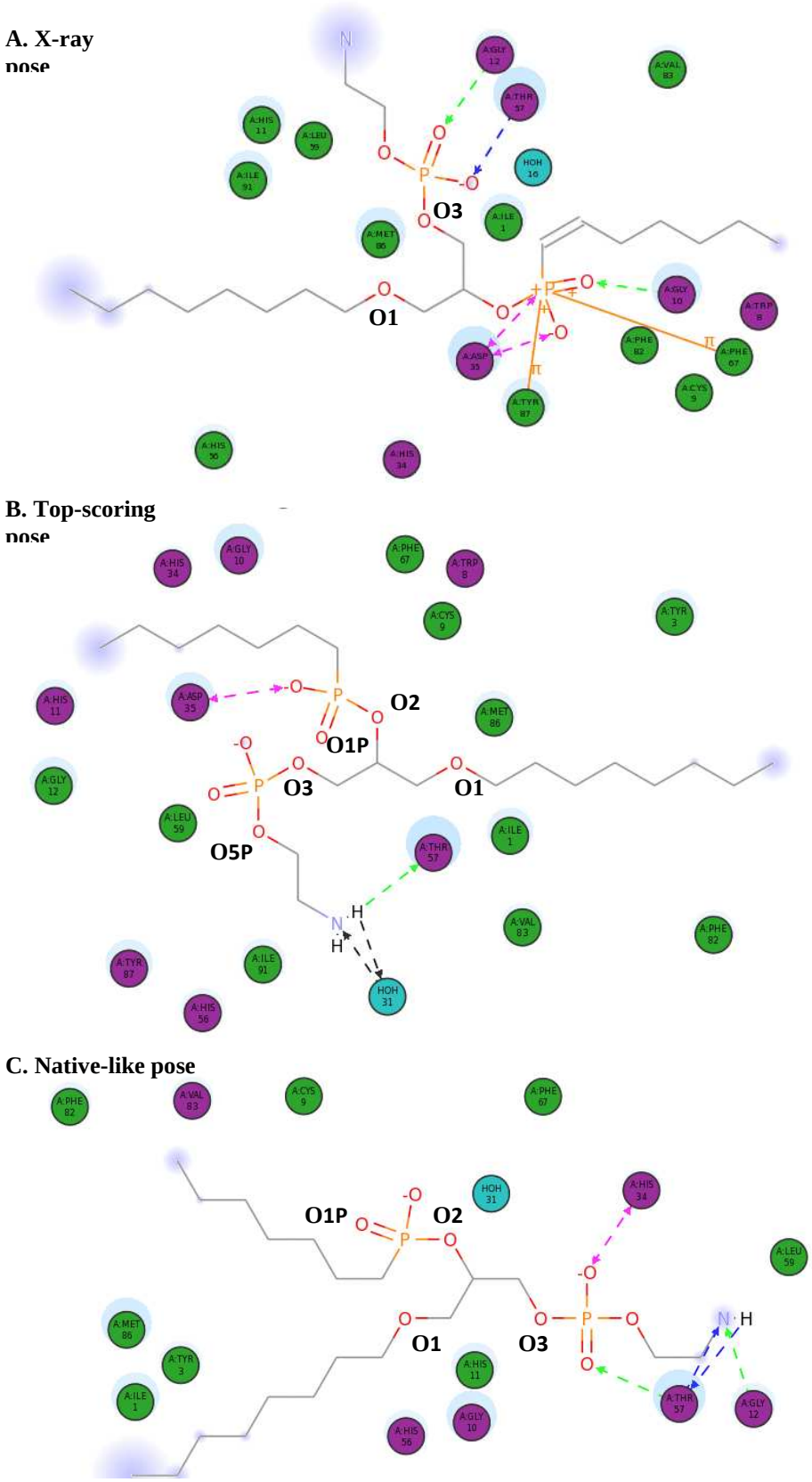


Figure 2.10. Hydrogen bonding interactions of GEL420 in 1poc. A). Crystallographic binding mode B). Top-scoring candidate pose C). Lowest RMSD pose. (blue/green/black: H-bond interactions, magenta: charge-charge interactions, dark blue circle: solvent accessible)

The results obtained from this study, therefore, suggest that poor ranking in the docking calculations can not be reliably associated with unsatisfied ligand donors and acceptors. Prioritising the results on unsatisfied hydrogen bonds would not improve the success rate of the docking.

Our analysis of the predictive power of lost hydrogen bonds for the assessment of poor docking poses should be contrasted with the more extensive incorporation of energy terms that reflect hydrogen bonding geometry and hydrophobicity as developed for the scoring function HYDE[117]. According to the scoring function, dehydration of non-polar groups contributes favourably to the overall score whereas dehydration penalties are associated with the burial of polar groups unless they are involved in a hydrogen bond with good geometry. The results reported for that program show that for some systems (estrogen receptor), such detailed considerations can improve the success of docking calculations, but that for other systems it has little effect (thrombin) or gives worse results (p38 kinase). The analysis presented here shows that across a wider range of targets, there are as many unmade hydrogen bonds in the "correct" X-ray pose as generated in the docking poses and this simplistic count of missing interactions is therefore not sufficiently discriminatory to allow incorrect poses to be identified.

Chapter 3

Weak Interactions in Protein-ligand Complexes: A survey of ligand aromatic ring acceptors

3.1. Introduction

3.1.1. Weak Hydrogen Bonds in Small molecules and Proteins

As described in Chapter 1, a classical hydrogen bond is represented as X–H…A, where both X and A are electronegative atoms. However, the ability of making hydrogen bonds is not just restricted to highly electronegative atoms such as F, N and O. In fact, a wide variety of hydrogen bonds exist involving weak donors such as acidic CH groups (e.g., CH groups flanked by strongly electronegative atoms) and weak acceptors such as π rings. Due to the varying character of donors and acceptors, hydrogen bonds span a wide energy range from 0.5 kcal mol^{-1} to 30 kcal mol^{-1} [137].

Weak hydrogen bonds involving aromatic rings and donors such as NH and OH groups have been well-studied in organic structural chemistry[138]. The significance of such interactions in the context of biological molecules was appreciated first in 1980's after observation of NH – π interactions in bovine pancreatic trypsin inhibitor and haemoglobin[139, 140]. In both studies an NH group was observed to point to the centre of an aromatic ring belonging to an adjacent side-chain or bound drug molecule. With the accumulation of high resolution protein crystal structures, systematic surveys of WHBs have been carried out to highlight distributions of geometric parameters associated with them and, in general, a mean distance between the donor atom and the ring centroid of 3.2 to 3.8Å was observed[25, 137, 141,

[142]. The role of WHBs as an additional stabilization factor in protein secondary structures has been proposed based on their frequent occurrence at helix termini[143] and β-sheets[144].

Studies from small molecules in the field of structural organic chemistry and crystal engineering have shed some light on the characteristics of WHBs[25]. For example, it was shown that highly activated CH groups have been shown to form hydrogen bonds with the strength approaching that of classical hydrogen bonds (about 7-8 kcal mol^{-1})[145]. However, as the acidity of the CH group decreases the significance of the weak interaction becomes negligible providing very little stabilization (about 0.5 kcal mol^{-1})[25]. Nevertheless, the interaction of CH groups with acceptors of varying strength (O, N, halogens and π rings) has been observed in small molecules and protein structures[134] and implicated in wide variety of phenomena such as crystal packing, supramolecular assembly[146] and macromolecular recognition[147].

Aromatic ring systems are involved in a variety of interactions in chemical and biological systems[23]. WHBs involving π rings as acceptors and donors of various strengths are well known. The analyses of non-bonded contacts observed in structural databases CSD[148] and PDB[149] have been routinely performed to study statistically favoured types and orientation of XH-π interactions. In the most recent survey of CSD, Bissantz *et al.* have noted that in CSD there is a clear preference for activated CH groups to interact with π rings in an above-ring orientation, with CH group pointing to the centre of the ring[10]. Similarly in protein systems, the aromatic side-chains, Trp, Tyr and Phe frequently interact with polarized CH groups[150]. Relatively stronger donors such as OH and NH interact slightly less frequently with

aromatic acceptors which is probably due to the attached desolvation cost of forming a hydrogen bond between a strong donor and π ring[10, 25].

3.1.2. Geometric and Energetic Considerations

The question pertaining to the significance of weak interactions such as CH…O and NH…π hydrogen bonds, particularly their energetic consequences on protein-ligand binding, has been a subject of debate[10, 25, 134]. Model systems such as benzene-water, benzene-ammonia and benzene-formamide complexes are used in theoretical investigations of weak hydrogen bonds[23]. The stabilization enthalpies of benzene H-bonding with ammonia and water, based on high-level CCSD(T) calculations, were estimated to be -2.0 kcal mol^{-1} (O…ring centre distance 3.4Å) and -1.61 kcal mol^{-1} (O…ring centre distance 3.6Å), respectively[151]. The experimental interaction energy value for benzene-water and benzene-ammonia weak hydrogen bonding was calculated to be -2.44±0.09 kcal mol^{-1} [152] and -1.84±0.12 kcal mol^{-1} [153], respectively, indicating close agreement between the two approaches[23]. The main contribution towards binding energy comes from long range interactions such as electrostatic and dispersion interactions[151]. This modestly stabilizing contribution is consistent with the prevalence of these interactions in secondary structures[134, 141, 154, 155].

Another study involving MP2 calculations on benzene-formamide complex, mimicking aromatic-amide interactions in proteins indicated interaction energy of up to -4.0 kcal mol^{-1} with two isoenergetic orientations [156]: T-shaped geometry, where XH vector (X=N/O) is perpendicular to the ring plane (Figure 3.1A), and alternatively parallel geometry, where XH is stacked above the ring[23, 156] (Figure 3.1B). The parallel stacking geometry between amide group and π rings has been observed to occur

frequently in peptides and proteins and can be rationalized on the basis that the amide group can still form additional hydrogen bonds whilst stacked on top of the ring, thereby giving additional stabilization[23, 137].

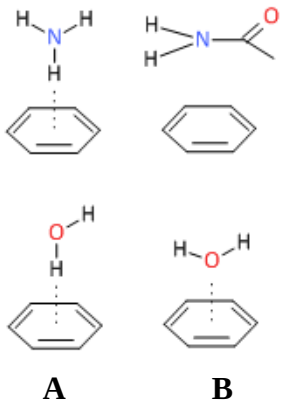


Figure 3.1. Favourable geometries of XH-π interactions. A. T-shaped geometry B. parallel stacking geometry.

Weak hydrogen bonds have been observed in protein-ligand complexes with pharmacological significance. Some of the major examples were surveyed by Toth *et al.*[134] where they highlighted the importance of these interactions in drug design. For instance, the difference in the binding affinity of different ligands for Protein Kinase C (PKC) δ was explained on the basis of additional CH-π interactions of higher affinity ligands in the binding site[157]. Similarly, the effect of substitution on a ring system in terms of its affinity to the target receptor was investigated by Schoepfer *et al.*[158]. A series of inhibitors of Src homology 2 (SH2) domain complexed with growth factor-bound receptor protein 2 (Grb2) showed different activities and the addition of electron donating groups on the indolyl moiety increased the affinity.

In some cases, the direct influence of weak hydrogen bonds on the ligand affinity is hard to determine and could be over-estimated, as argued by Bissantz *et al.*[10]. The case of ChK1 kinase ligands was highlighted for which it has been shown that the

favourable substitution on the phenyl ring leads to almost 100 fold increase in the affinity[159, 160]. They argue that the main factor behind affinity gain is not the NH-π interaction on its own, instead multiple interactions of the substituted ring.

Although it is not clear if weak hydrogen bonds at protein-ligand interface can lead to substantial gain in affinity, their supportive role in ligand design and lead optimization is well-established [10, 25, 134].

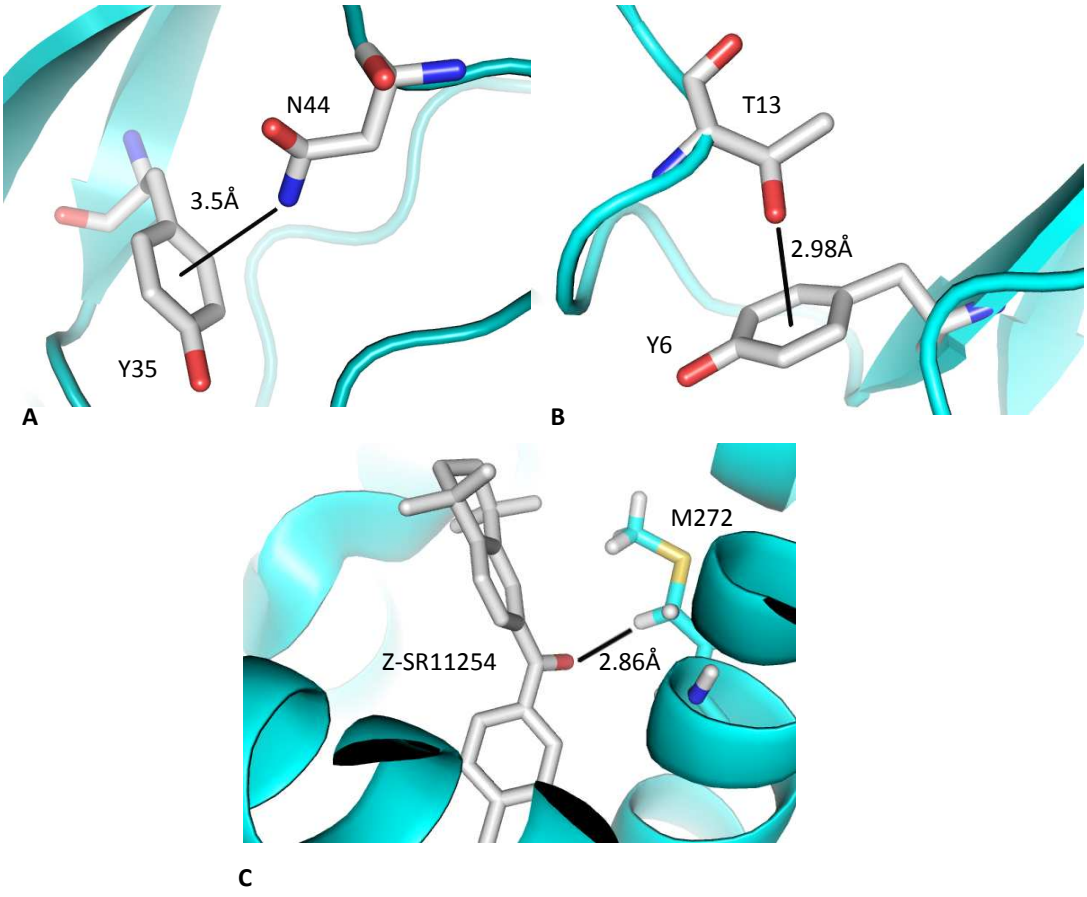


Figure 3.2. Examples of weak hydrogen bonds. A. NH-π interaction in bovine pancreatic trypsin inhibitor. B. OH-π interaction in glutathione transferase. C. CH-O interaction between retinoic acid receptor γ and Z-SR11254.

3.2. Aims

In the most recent survey of weak hydrogen bonds in protein-ligand complexes, particularly CH-π and XH-π (X= N, O) interactions, the geometric parameters and

contact densities of query atoms such as C, N, O around an aromatic ring were described[10]. The data originating largely from CSD[148] indicated that the perpendicular T-shaped geometry of CH-π interactions is observed more frequently when the CH group is attached to one or two hetero-atoms (N or O)[10]. This preference was however less well defined for NH and OH interactions with π-rings.

In this survey, we analyze the interaction preferences of CH, NH and OH groups in protein binding sites to ligand aromatic rings. We take into account the influence of ring substitution, fusion or presence of hetero-atom on geometric features and frequency distributions of contacts. The brief objectives of the study are:

1. Survey ligand ring systems in PDB and study frequency distribution of geometric parameters.
2. Identify any variation in the geometric pattern upon change in the type of ring system.
3. Compare the interaction geometries of strong donors (XH, X = N, O, S) and weak donors (CH) with π rings.

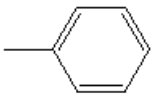
The distributions of CH, NH and OH groups around π-rings are expected to vary depending upon the character of the ring, for instance the nature of attached substituent, presence of heteroatom and fusion with another aromatic ring. It is generally accepted that an optimum geometry favours orientation of donor group pointing to the ring centre[23]. We therefore expect that changes in ring types such as the presence of an electron donating substituent would increase the frequency of observing optimum geometry or presence of an electron withdrawing substituent would decrease such occurrences.

3.3. Methods

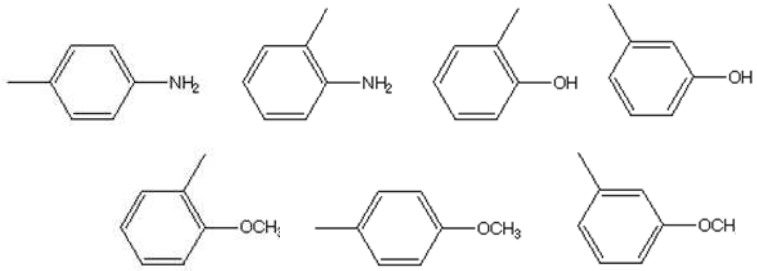
3.3.1. Dataset

The data for non-bonded contact analysis was obtained from IsoStar. IsoStar[14] is a knowledge-base that stores information about interactions between different chemical groups. The information in IsoStar is organized as pairs of interacting groups and their interaction could be shown as scatter plot based on all the instances of that interaction observed in CSD or PDB. One group in the pair is chosen as the central group and the other is called contact group. A contact is defined when the central and contact atoms are at distance which is sum of their van der Waals radii plus a 0.5Å tolerance. To generate the scatter plot the central group is least-squared superimposed and the spatial distribution of the contact group is displayed. Such plot highlights important features of the interactions such as directionality, orientation and geometrical preference of the contact group. Other features include statistical measures for a given interaction and theoretical calculations of interaction energies. IsoStar also supports querying interactions other than those already specified in the library using ConQuest which is available in its commercial package[14]. In this survey, the public IsoStar package was used to obtain spatial distribution of central and contact groups. The central groups (ligand π-rings) were divided into five categories and for each category, two contact groups (donors in protein binding sites) were analysed. Figure 3.1 shows further description of central groups. The five categories include, terminal phenyl group (Figure 3.3A), mono-substituted terminal phenyl with an electron donating functional group at one of three rings positions indicated (ortho-, para- and meta- positions) (Figure 3.3B), referred to as Phenyl-ED

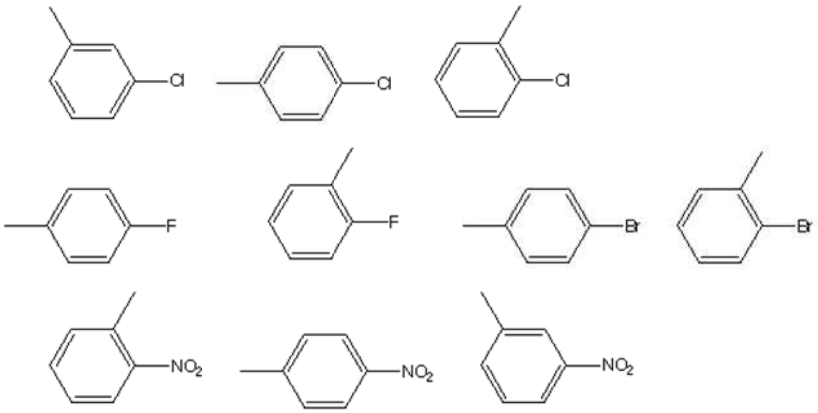
A. Phenyl



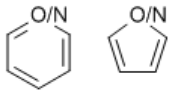
B. Phenyl with an electron donating substituent (Phenyl-ED)



C. Phenyl with an electron withdrawing substituent (Phenyl-EW)



D. 5 or 6 membered aromatic rings with a heteroatom (O or N)



E. Phenyl ring fused with another aromatic ring

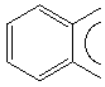


Figure 3.3. Ring types used to query IsoStar database in this study.

from this point, mono-substituted terminal phenyl with an electron withdrawing functional group at one of three (Figure 3.3C), referred to as Phenyl-EW from this point, phenyl fused to another aromatic ring (Figure 3.1D) and 5 or 6 member heterocyclic rings (with N, O as heteroatoms) (Figure 3.1E). Only mono-substituted phenyl rings were considered to study the effect of ring substitution, as with increasing number of substituents, additional interactions influence ring properties and it is hard to assign a category to the aromatic ring. Electron donating (ED) functional groups considered in this study include NH_2, OH, OCH_3 and electron withdrawing (EW) functional groups include NO_2, and halogens.

The contact groups were divided into two categories: weak donors (CH groups) and strong donors (XH groups where X = N, O or S). Although it is known that the donor strength is affected by flanking atoms but in this study, atoms connected to CH or XH groups are not considered and each group is treated as a potential donor.

The spatial distribution data were collected for each category of central group against each category of contact groups. Each file contained the aromatic ring and the atoms in contact, where threshold for contact distance was limited to sum of the van der Waals radii of the two atoms + 0.5Å. Any atom that is within this distance of any of the ring atoms is included in the analysis. Similarly where a substituent or a heteroatom in the ring is present, contacts to these atoms are also considered.

The resulting files were saved as SYBYL mol2 format for further analysis.

3.3.2. Non-bonded contact analysis

After obtaining contact atom distributions around central groups, each data file was processed with analysis scripts to extract and plot distributions of various geometric parameters. MolKit[121] package was used to read and analyse mol2 files. The location

of hydrogen atoms in protein structures is relatively uncertain, therefore, only heavy atoms were considered in this study.

For the aromatic ring, a centroid position was assigned and a normal vector from ring plane (M) was drawn (Figure 3.4). The distance of each contact atom (r) and the angle between M and vector pointing from contact atom to ring centre (w) was calculated. The distributions of the values for r and w parameters were obtained in each category. The in-plane distance of each contact atom from the ring centre (s) and its height from the ring plane (h) was also calculated to plot radial distribution of contact groups around the ring. The method for calculating radial distribution was based on Bissantz *et al.*[10]

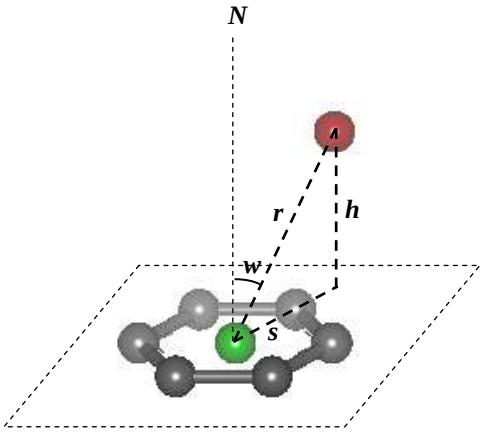


Figure 3.4. Geometric parameters calculated for weak interactions involving an aromatic ring system and a donor atom (O, N, S, C). An example of ring (grey) with its centre (green) and an interacting atom X (red) is shown. N: normal to ring plane, w: angle between N and vector pointing from X to ring centre, r: distance between X and the ring centre, h, height of X from ring plane, s, in-plane distance of X from the ring centre.

A 2-D grid of s and h values was constructed with 0.1Å x 0.1Å spacing, covering 7Å^2 with the ring centre at its origin. The number of contact atom counts at each grid point was calculated and scaled by the factor, F:

$F = N \times \frac{1}{2\pi r^2}$, where *N* is the total number of counts in the category and *r* is distance from ring centre.

The scaled values of contacts at each grid point were then plotted on an s versus h graph and each point was coloured (greyscale) where depth of the colour is proportional to the number of counts observed at that point. This graph simplifies 3-D spatial distribution plot into a simple 2-D plot where the propensity of contact atoms at a particular position from the ring can be observed.

3.4. Results

3.4.1. Frequency Distribution of Contacts

Table 3.1 summarizes the number of CH and XH contacts obtained from IsoStar for each category of ring system. The largest number of contacts was observed for heterocyclic rings, whereas the lowest number of contacts was observed for fused ring systems. In order to simplify the analysis, the criteria set for fused ring system was a phenyl ring attached to two aromatic carbon atoms which could eliminate some of the more complicated ring types. The number of CH and XH contacts decrease successively for phenyl, phenyl-ED and phenyl-EW categories.

From the total number of contacts found in each category, radial distribution plots were generated to show contact densities of query atoms above the rings (Figure 3.5). The frequency of observing a donor atom at a particular position around the ring is projected to a 2-D plot whose coordinates are defined by the in-plane

distance from ring centroid (s) and height above the ring plane (h). The intensity of color describes the frequency with dark color representing high frequency. The origin of the plot can be considered as the ring centroid. As the distance between ring centre and one of the ring atoms is about 1.5 Å therefore s values within 1.5 Å represent interaction above the ring.

Table 3.1. Number of CH and XH (X= N, O or S) contacts found in IsoStar for each of the ring system category studied in this survey.

Ring System	CH contacts	XH contacts
Phenyl	2598	1655
Phenyl with an electron donating substituent	2171	1184
Phenyl with an electron withdrawing substituent	986	480
Phenyl ring fused another aromatic ring	509	282
Heterocyclic ring (5/6 members)	5066	5020

The contacts between the donor atom and the aromatic rings can be defined by s values in the range of 4.5 Å to 5.5 Å and h values in the range of 3.0 Å to 4.0 Å. It is to be noted that for Phenyl-ED and Phenyl-EW category, contacts outside these values of s and h could possibly include contacts to the substituent atoms.

For phenyl rings, the distribution of CH groups is almost uniform around the ring and concentrated primarily within the above-mentioned ranges of s and h values, 4.5-5.5 Å and 3.0-4.0 Å, respectively (Figure 3.5A). The distribution of XH groups is roughly similar but with two subtle differences. Firstly, the frequency of contacts is slightly lower above the ring (s = 0.0-1.5Å) as compared to other regions.

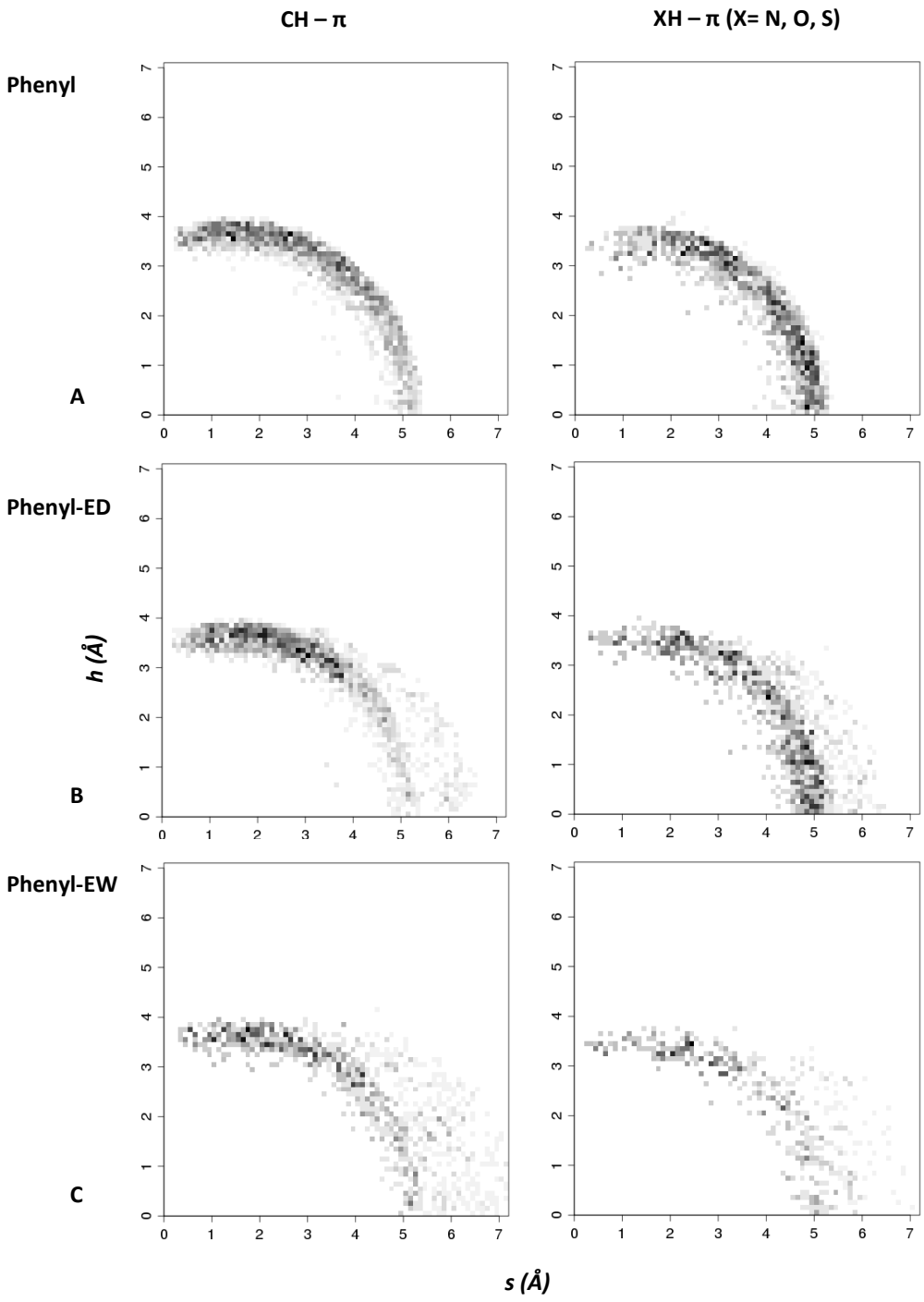


Figure 3.5. Radial distribution plots of CH and XH (N, O or S) atoms around different aromatic ring systems in the PDB. The centre of the ring coincides with the origin of the plot and x-axis (*s*) is the in-plane distance of contacts atoms from the ring centre and y-axis (*h*) is the height of contact atoms above the ring plane. The frequency of contacts is proportional to the darkness of colour, at each point in the plot.

This was unexpected as strength of XH donors should dictate a preference towards above-ring orientation. Secondly, the lower limits of XH contacts, both in terms of s and h are not as well defined as for CH groups. This indicates that shorter contact distances can be observed for strong donors (Figure 3.5A).

The distribution of CH groups around phenyl-ED rings shows a clear preference for T-shaped geometry whereas it diminishes around the ring (Figure 3.5B). It was mentioned in Methods section (3.3.1) that contacts to substituent atoms are also included which can be observed in this case by the spread of distribution at large s and small h values. The distribution for XH…phenyl-ED contacts shows a different pattern. It can be observed that contact frequency above the ring is lower than what was observed for CH…phenyl-ED contacts and XH…phenyl contacts. The contacts distribution around the ring shows a wide range of geometric orientations. This can be explained by the presence of contacts to substituent atoms (Figure 3.5B).

Finally, contact distributions around phenyl-EW rings show lowest frequencies among other ring types (Figure 3.5C). This observation is consistent with the idea that an electron withdrawing functional group should decrease a π ring's capacity as an electron acceptor. For CH groups, the T-shaped geometry appears to be most frequently encountered, among the observed contacts. For XH groups a clear pattern can not be observed due to a very small number of contacts observed which appear to be randomly distributed (Figure 3.5C).

Figure 3.6 shows radial distribution plots of CH and XH contacts with fused ring systems. It was expected that the presence of a fused ring could provide a large contact surface area therefore more chances for contacts above the ring. In general, a high frequency at an s-value of 0.0-1.5Å should indicate a preference towards

above-ring orientation. In the case of fused rings, a relatively high frequency region appears to be spread over s values of 0.0-2.5Å (Figure 3.6). This could arise from contacts over an extended region above the ring provided by the fused rings system. Similar, observation was made for XH groups although in general a much lower overall frequency of contacts was also noticed (Figure 3.6).

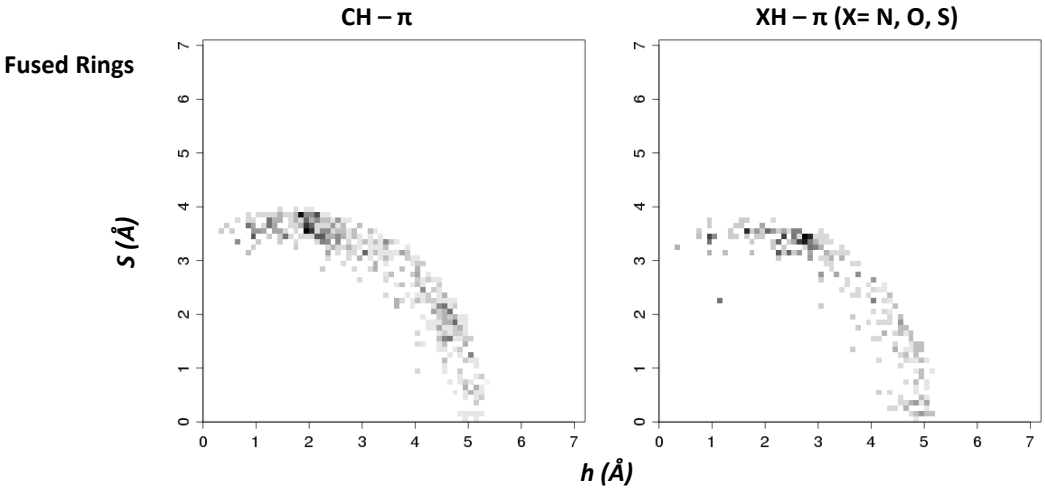


Figure 3.6. Radial distribution plots of CH and XH (N, O or S) atoms around fused ring systems in the PDB.

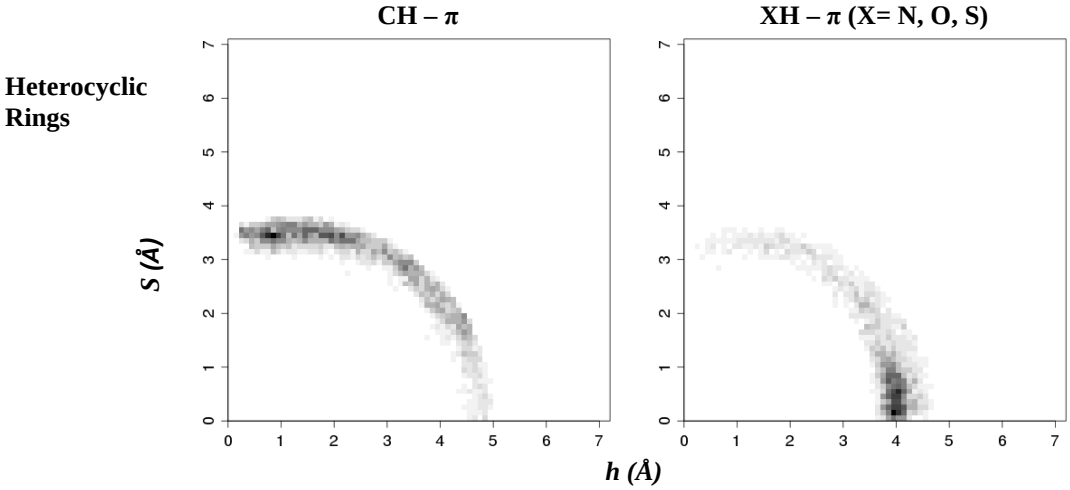


Figure 3.7. Radial distribution plots of CH and XH (N, O or S) atoms around heterocyclic ring systems in the PDB.

The last category of aromatic rings includes 5- or 6-membered heteroaromatic ring systems. For CH groups contact density is relatively higher above the ring but gradually diminished while moving away from the ring (Figure 3.7). This trend is completely revered for XH groups which indicates the when a strong donor is

present interactions to hetero-atoms, which most probably involve hydrogen bonding, is clearly preferred (Figure 3.7).

3.4.2. Distribution of Geometric Parameters

In proteins, aromatic-amide interactions have been observed at donor-centroid distance of ≥ 3.5 Å[10]. The distribution of donor-centroid distance (r) and the donor angle from the normal to ring plane (w) was compared among different classes of ring types. The resulting histograms are shown in Figure 3.8.

The vast majority of both CH-π and XH-π contacts is observed at r values higher than 3.5Å. For phenyl rings, CH contacts have a sharp peak at 4.8Å whereas for phenyl-ED rings, there is a relatively flat peak at 4.5 to 5.0Å (Figure 3.6A). There is an additional peak at much longer donor-centroid distance. These contacts could possibly include interactions with substituent atoms and are not discussed in the current comparison. CH contacts with phenyl-EW rings do not show a distribution with as well-defined peaks as observed for phenyl and phenyl-ED rings. However, the highest frequency is observed at around 5.0Å. The spread of the distribution is also wider than other two categories of rings. It was expected for phenyl-EW rings to have an almost random distribution of contacts with donor groups as their strength as acceptor is much lower than phenyl or phenyl-ED rings (Figure 3.6A).

For XH contacts the shape of r distributions did not show much difference among phenyl, phenyl-ED and phenyl-EW rings (Figure 3.6A). For phenyl rings, XH contacts show a sharp peak at 4.8Å. For phenyl-ED rings, there was similar peak but the distribution is spread further towards longer distances. Finally, for phenyl-EW rings, the shape of XH contact distribution is very similar to what was observed for CH…phenyl-EW contacts (Figure 3.6A).

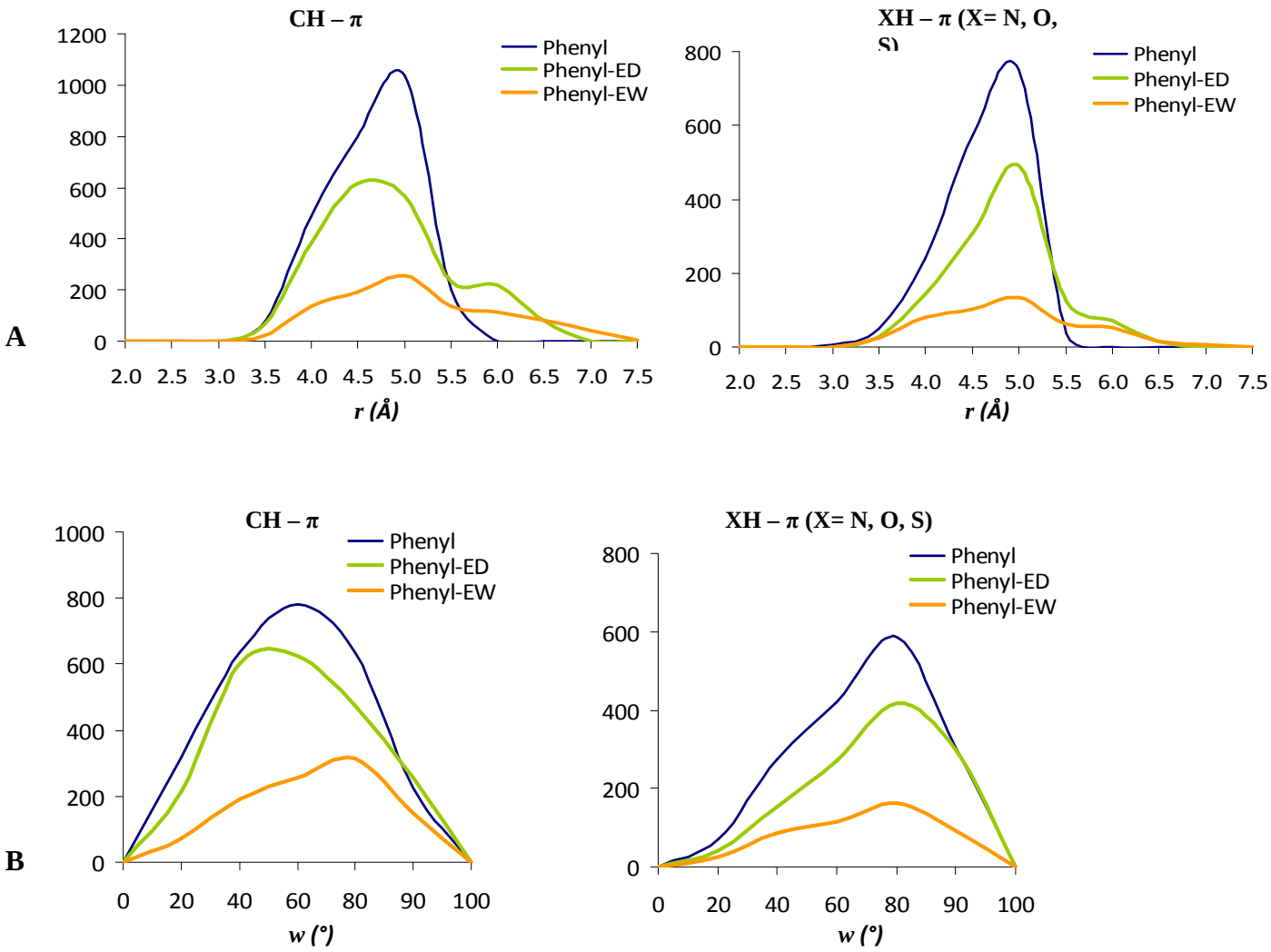


Figure 3.8. Distributions of geometric parameters in CH…π and XH…π contacts. A, *r*: distance between donor atom and ring centre (d). B, *w*: angle between normal to ring plane and the vector pointing from donor atom to ring centre.

The angular distributions for CH and XH contacts with different ring types are shown in Figure 3.6B. As observed for the distance distributions, the shape of angular distributions for CH contacts varies more significantly with different ring types than for XH contacts (Figure 3.6B).

CH contacts with phenyl rings show a symmetrical distribution centred around 60°. For phenyl-ED rings, the distribution is skewed towards shorter angle with peak observed at 40°. For phenyl-EW, the opposite trend is observed with the highest

peak at 80°. Interestingly, XH contacts did not show much difference in the shape of angular distribution with almost all three distributions showing a peak at 80° (Figure 3.6B).

The median values of distance and angle distributions are reported in Table 3.2. As the mean values can be affected by the outliers such as contacts to substituent atoms, therefore, median values represent a better assessment of the distributions.

Table 3.2. Median values for distance (d) and angle (w) distributions of CH and XH contacts with ring systems in the PDB.

Ring	CH		XH	
	d (Å)	w (°)	d (Å)	w (°)
Phenyl	4.5	49.5	4.5	62.7
Phenyl-ED	4.6	48.0	4.7	67.8
Phenyl-EW	4.8	57.0	4.7	61.7

3.5. Discussion

Theoretical studies and database surveys of interactions involving aromatic rings and chemical moieties such as CH, OH and NH groups have pointed towards two isoenergetic orientations[23, 134]. In a T-shaped geometry, the XH group is located above the ring, pointing towards its centre (Figure 3.1A). In a stacking geometry, the XH group is also located above the ring however it is parallel to ring plane (Figure 3.1B). Other less favourable configurations in which XH group is located around the ring and in some case interacting with one of the –CH groups in the ring are also possible[151, 156]. In a survey of aromatic-amide interactions in 592 high resolution protein structures (≤ 1.6Å) stacking geometry was observed more frequently[137]. In another recent survey of molecular interactions, CH-π contacts involving polarized CH groups (such as those flanked by heteroatoms, O/N) were shown to have a

preference towards an above-ring orientation (C on top of the ring and CH vector pointing to ring centre)[10].

In this study, the above ring orientation includes both T-shaped and stacking geometries as the orientation of CH vector relative to ring centre was not considered.

It is expected that the interaction preferences of aromatic rings should be affected by the nature of attached substituents. A study on a series of Grb2-SH2 inhibitors indicated that the type of substituent on an indolyl group which was involved in a CH-π interaction influenced the affinity of the inhibitor. For example, the increase in affinity was related to the increasing electron donating character of the substituent[134, 158]. In another *ab initio* fragment molecular orbital study on a set of leucocyte-specific protein tyrosine (LCK) kinase inhibitors, the role of CH-π and NH-π hydrogen bonds was highlighted[161]. Moreover, a ten-fold increase in the affinity of an inhibitor was attributed to the modulation of CH-π and NH-π interactions which was achieved by replacing an electron withdrawing chloro- functional group with two electron donating methyl groups on an aniline ring in the ligand[161].

Based on this idea, a survey of ligand aromatic rings in protein-ligand complexes from PDB was conducted to investigate the orientation of CH and XH contacts (where X = N, O, S) around different types of rings (Figure 3.3). A qualitative assessment of interaction preferences of rings with different substituents was done by plotting the distribution of geometric features of aromatic interactions (Figure 3.4). Due to ambiguity in the position of hydrogen atoms in protein structures, it was not possible to discriminate between T-shaped and stacking geometry. A distinction could, however, be made between an above-ring orientation (which include both T-

shaped and stacking geometry) and other less favourable configurations around the ring. It was expected that the above-ring interaction should be more favourable for XH groups than for CH groups. Similarly, electron donating substituents and fused rings with larger contact surface area should also favour above-ring orientation.

The results of the survey indicated some interesting trends. One of the most important aspects was that the variation in interaction preferences was more well-defined for CH-π contacts then for XH-π contacts (Figure 3.5 and 3.8). For example, the above-ring geometry preference in the case of CH…phenyl-ED contacts was relatively more obvious than for XH…phenyl-ED contacts (Figure 3.5). Similarly, the change in the shape of distance and angular distribution was more visible for CH groups (Figure 3.8). These results are similar to the trend noticed in a recent database survey which showed more clear preference of polarized CH groups for above-ring geometry than for OH/NH groups[10]. This trend was explained by the increasing strength of OH/NH-π interactions which could in principle impose a higher desolvation cost on binding energy. Therefore, in general it is expected for ligands to employ strong donors in hydrogen bonding with strong acceptors to overcome desolvation penalty and achieve optimal binding. It can therefore be expected XH-π weak hydrogen bonds should be formed with most optimum geometry. We have observed that for above-ring orientation, which corresponds to in-plane distance (s) within 0.0-1.5Å, the minimum above-plane height (h) for XH groups (2.7Å) was shorter than that of CH groups (3.0Å) (Figure 3.5A). One such example is shown in Figure 3.9A. This example is based on the complex between third PDZ domain of synaptic protein and its peptide ligand. A Phe ring in the ligand is shown to form a hydrogen bond with NH group of an Asn side-chain.

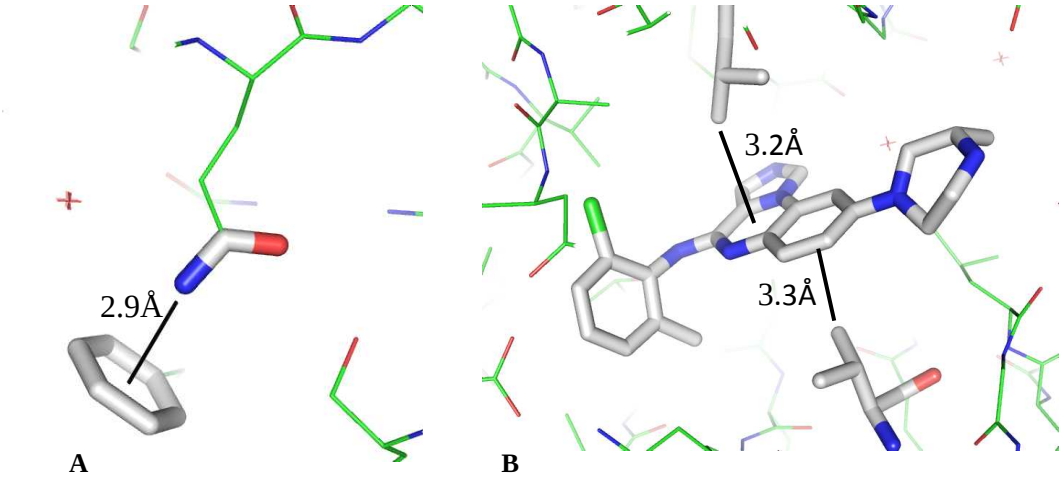


Figure 3.9. Selected examples of weak interactions from IsoStar. A. An NH-π interaction in PSD-95 (PDB code: 2KA9) at a very short donor-centroid distance B. An example of multiple CH-π interactions with fused ring systems observed in LCK kinase (PDB code: 3AD4).

The idea that XH groups preferably form hydrogen bonds with strong acceptors is further strengthened by contact distributions of XH groups around heterocyclic rings (Figure 3.7). Almost opposing trends were observed for CH and XH groups. CH groups show wide range of configuration around heterocyclic rings with relatively less contact density around the hetero-atom. XH groups on the other hand were predominantly involved in interactions with the heteroatom. The heteroatom in the ring is a potential acceptor and an unsatisfied ligand acceptor in protein binding site is extremely unfavourable[6], as noted in the survey of protein-ligand complexes in Chapter 2.

Another consideration in aromatic interactions is the contact surface area of the π ring. In a survey of 592 high resolution protein structures, 17.5% of Trp side-chains were involved in weak hydrogen bond which was almost twice the percentage of other aromatic residues such as Tyr (8.8%) and Phe (5.8%)[137]. This was rationalized

on the basis of larger aromatic surface and conjugation of fused ring systems which could possibly increase the acceptor strength. In our analysis of fused ring systems in ligands, an above ring preference appeared to span a broader range of s values (0.0 to 3.0Å) (Figure 3.6). As the fused ring could also act as second acceptor so the broad range of s values reflects interaction above the fused ring systems. The relatively higher frequency is therefore consistent with the stronger interaction of CH or XH groups with fused rings. One such example related to ligand binding in LCK kinase is shown in Figure 3.9B. This represents a reported case of CH interactions with fused ring in the ligands and with role in binding affinity[161]. In fact a ten-fold increase in the affinity was noticed when this CH-π interaction involved aromatic rings with electron donating substituents.

Surveys of molecular interactions in X-ray crystallographic structures always have some limitations[6, 10]. For example, the resolution of the majority of structures lie between 2 to 3 Å which corresponds to a standard deviation in the position of atomic coordinates of up to 0.4Å[6]. This is further complicated by uncertainty in the position of hydrogen atoms, indistinguishable side-chain orientations such as His, Asn, Thr and mobile water molecules.[6, 19] Furthermore, protein-ligand binding is a dynamic phenomenon where both protein and ligand could exhibit multiple bound states[35]. Therefore, the general trends observed in such surveys give only a qualitative assessment of the variation in interaction geometries with the nature of interacting moieties.

The trends observed in this survey are consistent with previous studies however they have been investigated in a different setting. For example, a recent survey of CH-π and XH-π weak interactions compared interaction geometries of polarized CH groups

to NH/OH groups. In this survey, we focused on the types of ligand aromatic rings and analyzed their interaction preferences. It is intuitively expected that an increase in electron density of the aromatic ring should stabilize its interaction with CH or XH groups. In a combined theoretical and spectroscopic study, Gosling *et al.* showed that both vibrational frequency and *ab initio* calculations indicate stabilization of 4-flurotoluene–ammonia complex formation[162].

This database survey further supports these results based on the frequent occurrence of favourable above-ring orientation of CH or XH groups (T-shaped and stacking geometry) as compared to other configurations around the ring. It was also noticed that the above-ring interaction preference was more pronounced for CH groups than for XH groups. This is an interesting observation as the role of CH and XH groups involved weak interactions and their comparison has been a subject of debate. An important question in this regard is if XH-O interactions are interchangeable with CH-O interactions and what consequence it should have on the binding affinity?"[25]. Pierce *et al.* investigated strong CH-O interactions in optimized kinase ligands and reported that the replacement of standard hydrogen bonds with their CH analogs resulted in almost similar binding affinities[160]. The comparable strengths of the two types of hydrogen bonds were explained on the basis of higher desolvation penalty on the part of strong donor groups. It was noted that in the cases where this kind of interchange does not bring about significant affinity changes there should be significant opportunities in optimizing non-binding related properties of inhibitors such as solubility, permeability and metabolism[25, 160].

Chapter 4

Predicting solvent and fragment positions in protein binding sites using MCSS and CHARMm

4.1 Introduction

As described in Chapter 1, fragment-based lead discovery has gained huge interest for its ability to efficiently sample chemical space and providing suitable starting points for drug discovery. In the field of computer-aided ligand design, equivalent techniques are under development for a long time and a wide range of methods now exist to probe binding sites for favourable positions of functional groups and small-molecule fragments. Some of these methods were reviewed in Chapter 1 including Multiple Copy Simultaneous Search (MCSS), one of the pioneering efforts in this area. The challenges in predicting fragment binding to proteins still remain such as appropriate treatment of solvent molecules, flexibility of the binding site and correct protonation states. The work described in this chapter investigates the performance of the current version of MCSS calculations as a method for probing binding sites for fragment binding. The results were compared with similar calculations using the program GOLD, performed by Hugues-Olivier Bertrand of Accelrys.

4.2 Aims

In this study, MCSS calculations were performed on different datasets to investigate docking and scoring of solvent molecules and fragments to protein binding sites. There were 5 main components to the study:

- The success rate of MCSS docking and scoring was evaluated for reproducing and correctly ranking experimentally observed positions of solvents and fragments in the protein binding sites of a dataset of experimental structures.
- A rescoring method using an implicit solvent model was assessed for its ability to improve the ranking of correct protein-fragment poses in the dataset.
- The same dataset was used to evaluate the performance of the docking program GOLD and the solvent mapping program, FT-Map.
- A preliminary assessment was made of the importance of protein flexibility by assessing the impact of using multiple protein structures in the MCSS calculations.
- The effect of including conserved water molecules in the protein binding site.

4.3 Datasets

4.3.1 Elastase Dataset

Porcine pancreatic elastase (elastase) is a serine protease with a characteristic catalytic triad in the binding site cleft consisting of Ser206, His60 and Asp108. Its well-characterized active site consists of multiple sub-sites that accommodate amino acid residue flanking the peptide bond that is to be cleaved. These sub-sites were named S1, S2, S3, S4/S5 and S1', S2', S3' before and after the scissile bond, respectively[78] (Figure 4.1). The experimental solvent mapping was performed by Mattos and co-workers using different solvent probes which showed at least 16 unique binding sites for different organic solvents (numbered from 1006 to 1010), 6

of which corresponded to the sub-sites in the binding cleft, whereas the rest were mostly located in crystal contacts[82].

In total 9 crystal structures from this study were available from the PDB. As we are interested in predicting solvent binding sites that lie in the active site cleft only the relevant structures were selected for the analysis. The PDB codes of these structures, names of the bound solvent, their experimentally observed binding sites and corresponding binding cleft sub-sites are shown in Table 4.1.

Table 4.1. Solvent-bound X-ray structures of Elastase used in this study. Solvent-binding sites that overlap with S1, S3, S4, S1' and S3' sub-sites are considered. For each structure, PDB code, resolution (R), name of bound solvent and occupied sub-sites are given.

PDB	R (Å)	Solvent	S1	S3	S4	S1'	S3'
2FO9	2.0	Acetone	ACN1001	-	-	-	-
2FOA	1.9	Isopropanol	IPA1001	-	IPA1002	-	-
2FOC	2.0	Dimethylformamide	-	-	-	-	DMF1004
2FOD	2.0	Ethanol	ETH1001	ETH1003	ETH1002	-	ETH1004
2FOE	2.2	5-Hexene-1,2-diol	HE X1001	-	-	-	HEX1004
2FOG	1.9	Trifluoroethanol	TFE1001	TFE1003	TFE1002	TFE1008	-

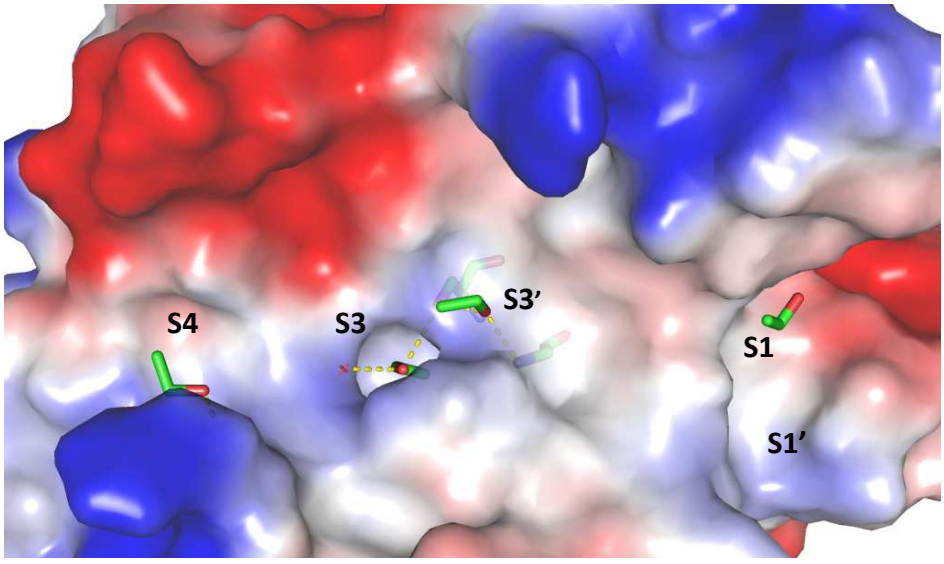


Figure 4.1. Binding site of elastase with experimentally determined positions of ethanol solvent probe.

4.3.2 Thermolysin Dataset

Thermolysin is a metalloproteinase that specifically cleaves peptides bonds containing hydrophobic residues. Its large active-site cleft contains four sub-sites, named S2, S1, S1' and S2', where S1' forms the main specificity pocket having preference for hydrophobic residues[80] (Figure 4.2). Experimental mapping studies were performed using isopropanol[80], acetone, acetonitrile and phenol[81].

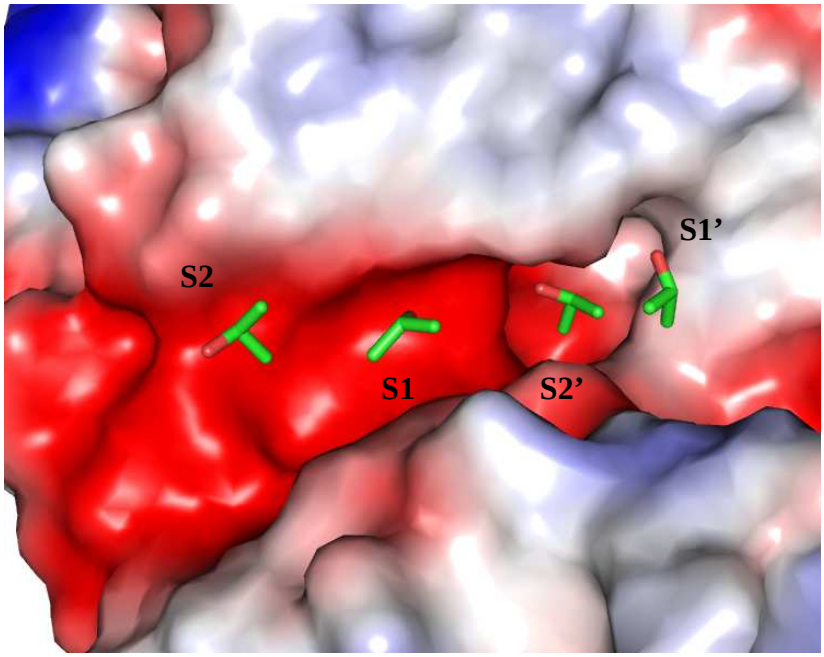


Figure 4.2. Binding site of thermolysin with experimentally determined positions of isopropanol solvent probe.

Table 4.2. Solvent-bound X-ray structures of Thermolysin used in this study. Only the solvent-binding sites in the active site are considered. For each structure, PDB code, resolution (R), names of bound solvent molecules and occupied sub-sites are given.

PDB	R (Å)	Solvent	S2	S1	S1'	S2'
1FJQ	1.7	Acetone	-	-	ACN1	-
1FJU	2.0	Acetonitrile	-	-	CCN1	-
1FJW	1.9	Phenol	-	-	IPH1	-
8TLI	2.2	Isopropanol	IPA5	IPA8	IPA1	IPA9

All four solvent molecules bind to the main specificity pocket S1' whereas other sub-sites were also occupied by at least one of the solvent probes. The PDB structures corresponding to solvent binding sites overlapping with the active site cleft were used in this study. Table 4.2 summarizes the details of these structures.

4.3.3 Fragment docking dataset

Recently, a set of 12 fragment-protein complexes was described which consists of targets investigated using fragment-based methods and for which X-ray structures bound to various fragments are available in the PDB[59]. This dataset will be referred to as the 'fragment docking dataset'. The protein targets, their corresponding PDB codes and literature references are listed in Table 4.3. The chemical structures of the fragments are shown in Table 4.4.

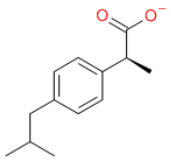
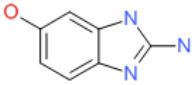
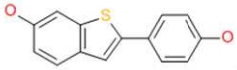
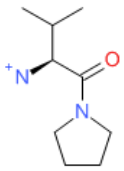
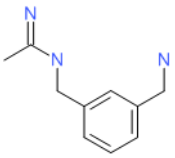
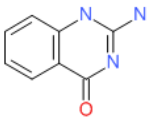
4.3.4 HSP90 dataset

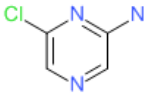
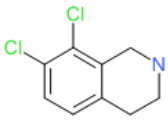
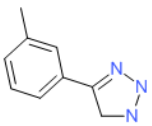
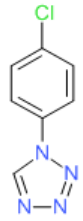
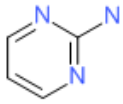
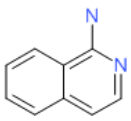
Heat shock protein (HSP) 90 is well-known for its function as a molecular chaperone. Due to its important role in assisting protein folding, preventing self-aggregation and cell cycle progression, it has been established as a valuable target in anti-cancer drug development. The structure of HSP90 contains a highly conserved N-terminal domain that is linked, *via* a highly flexible linker, to a middle domain and a C-terminal domain[163]. The N-terminal domain contains an adenosine binding pocket, responsible for its ATPase activity and has an unusual motif known as a Bergerat fold[163].

Table 4.3. List of protein-ligand complexes in fragment docking dataset.

PDB code	Receptor	Resolution (Å)	Reference
1EQG	Cyclooxygenase-1 (COX-1)	2.60	Selinsk *et al.* 2001 [164]
1FV9	Urokinase (uPA)	3.00	Hajduk *et al.* 2000 [165]
1GWQ	Estrogen Receptor α (ER)	2.45	Warnmark *et al.* 2002 [166]
1N1M	Dipeptidyl Peptidase IV (DPP-IV)	2.50	Rasmussen *et al.* 2003 [167]
1S39	tRNA Guanine Transglycosylase (TGT)	1.95	Meyer *et al.* 2004 [168]
1WCC	Cyclin Dependent Kinase 2 (CDK2)	2.20	Hartshorn *et al.* 1999 [169]
1YZ3	Phenylethanolamine N-Methyl	2.40	Wu *et al.* 2005 [170]
2ADU	Methionine Aminopeptidase (MetAp2)	1.90	Kallander *et al.* 2005 [171]
2C90	Thrombin	2.25	Howard *et al.* 2006 [172]
2JJC	Heat Shock Protein 90 (HSP90)	1.90	Congreve *et al.* 2008 [59]
2OHK	β-Secretase (BACE-1)	2.20	Murray *et al.* 2007 [173]

Table 4.4. List of fragments in fragment docking dataset. (RB: rotatable bonds, MW: molecular weight), rotatable bonds consider only rotation around single bond, hydrogens are omitted for simplicity.

PDB	Fragment	RB	MW
1EQG		4	206
1FV9		0	149
1GWQ		0	242
1N1m		2	171
1QWC		3	177
1S39		0	161

PDB	Fragment	RB	MW
1WWC		0	130
1YZ3		0	202
2ADU		0	161
2C90		0	181
2JJC		0	95
2OHK		0	144

This fold is characterized by a rigid adenosine binding site and a flexible loop in a phosphate binding region, acting mostly as an active-site lid[174]. The N-terminal domain of HSP90 is known to undergo conformational changes upon binding to various ligands[38]. The ATPase activity of the N-terminus drives structural transitions required for chaperone functioning therefore it has been targeted in several drug discovery campaigns[175].

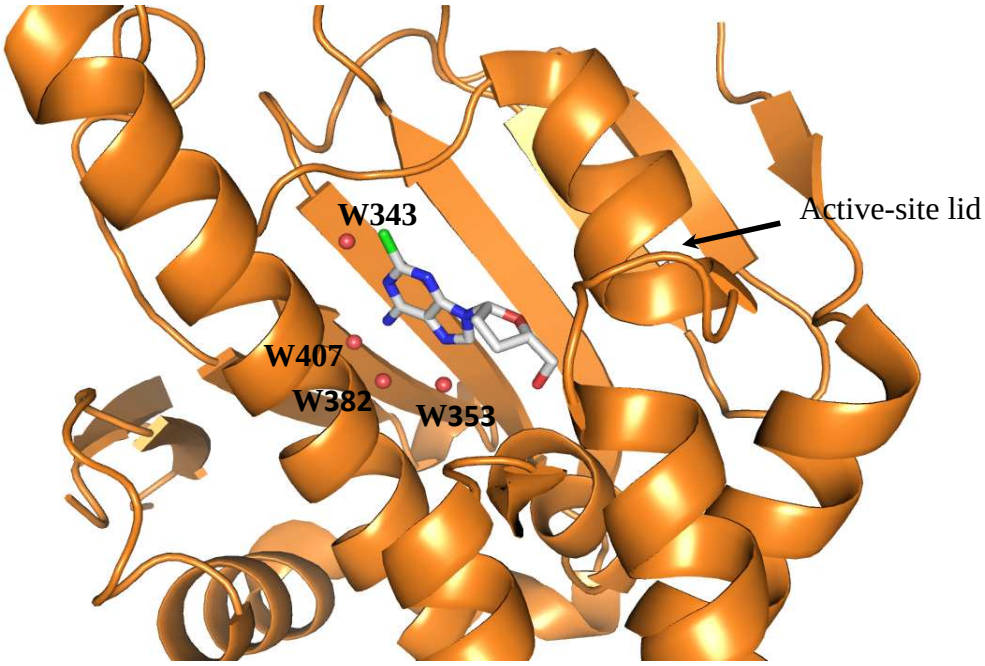


Figure 4.3. HSP90 N-terminal domain active site. Endoplasmic reticulum paralog of cellular HSP90 is shown with 2-chlorodideoxyadenosine inhibitor bound to the active site, along with 4 water molecules that are highly conserved in HSP90 protein-ligand complexes (PDB code: 1QYE).

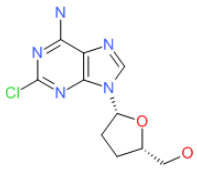
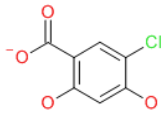
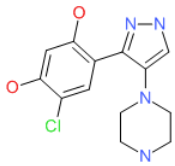
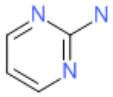
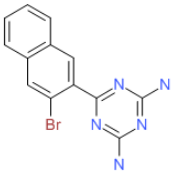
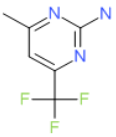
In order to study the performance of MCSS calculations in a protein-ligand docking context, a set of HSP90 structures bound to various fragments was included in this study (Table 4.5). There are 3 to 4 highly conserved water molecules in the active site of almost all HSP90 structures in the dataset (Figure 4.3). These water molecules

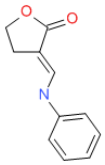
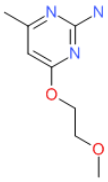
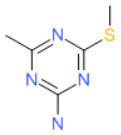
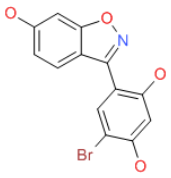
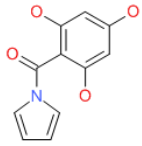
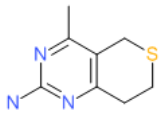
play an important role in ligand binding by bridging interactions between ligand and protein atoms, as described in the references mentioned in Table 4.5. The chemical structures of the fragments are shown in Table 4.6.

Table 4.5. List of HSP90-fragment complexes used in HSP90 dataset.

PDB code	Resolution (Å)	Reference
1QYE	2.10	Soldano*et al.* 2003 [176]
1ZWH	1.65	Immormino, R.M (*to be published*)
2CCS	1.79	Barril *et al.* 2006 [177]
2JJC	1.95	Congreve *et al.* 2008 [59]
2QF6	3.10	Huth *et al.* 2007 [43]
2QFO	1.68	Huth *et al.* 2007 [43]
2WI1	2.30	Brough *et al.* 2009 [44]
2WI2	2.09	Brough *et al.* 2009 [44]
3BM9	1.60	Gopalsamy *et al.* 2008 [178]
3EKO	1.55	Kung *et al.* 2008 [179]
3FT5	1.90	Barker *et al.* 2009 [180]

Table 4.6. List of fragments in HSP90 dataset. (RB: rotatable bonds, MW: molecular weight), rotatable bonds consider only rotation around single bond, hydrogens are omitted for simplicity.

PDB	Fragment	Res	RB	MW
1QYE		2.10	2	270
1ZWH		1.65	1	188
2CCS		1.79	2	295
2JJC		1.95	0	95
2QF6		3.10	1	316
2QFOa		1.68	1	177

PDB	Fragment	Res	RB	MW
2QFOb		1.68	0	189
2WI1		2.30	4	183
2WI2		2.09	1	156
3BM9		1.60	1	322
3EKO		1.55	1	219
3FT5		1.90	0	181

4.4. Methods

All the steps described below for preparation of input structures, docking and scoring were performed in Discovery Studio 2.5[181].

4.4.1 Preparation of Receptor Structures

The general steps for the preparation of receptor structures, applicable to all different datasets, are as follows. In order to prepare receptor structures for docking, all ligands and water molecules were removed, except for some cases that are described separately. Only one set of conformations was kept for side chains with alternate conformers. The resulting protein chains were assigned CHARMm atom types and MMFF94 partial charges. Hydrogen atoms were placed and their positions optimized with CHARMm energy minimization. Each receptor structure was subjected to 5000 cycles of energy minimization using Adopted Basis Newton-Raphson algorithm. Heavy atoms were fixed during the minimization and distance-dependent dielectric model was used for approximating the solvent. The calculations were performed with neutral histidine residues.

Elastase: For elastase, the binding sites for placement of solvent molecules was defined as a sphere around S1, S3, S4, S1' and S3' sub-sites, which is roughly centred on S3 sub-site (43.38, 24.06, 35.04) and has a radius of 17Å.

Thermolysin: The binding site for thermolysin was defined as a sphere around S2, S1, S1' and S2' sub-sites, whose centre is located in S2 sub-site (34.71, 41.11, -7.16) with a radius of 10Å. The Zn^{2+} ion in thermolysin binding site and water molecule(s) coordinated to it were not removed during the calculations.

Fragment docking and HSP90 datasets: In order to prepare receptor structures, all ligands and water molecules were removed, except for HSP90 (2JJC) and PNMT (1YZ3), where water molecules bridging interactions between the fragment and the protein binding site were kept. The conserved water molecules in the binding site of all structures in the HSP90 dataset were also kept (Names of these water molecules in PDB files are provided in Table 5.9). All ions were removed except for MetAp2 (2ADU) where two Cobalt (Co^{2+}) ions are involved in key interactions with the bound fragment. Where two or more copies of the protein-fragment complex were present in the asymmetric unit, the copy with the lowest B-factors of binding site residues was selected. Only one set of conformations was kept for side chains with alternate conformers (conformer A). The standard CHARMm atom typing and parameter set were used for the protein atoms. Histidine residues were treated as neutral. These include His residues interacting with the fragments (for 1GWQ and 2ADU) in which case it was confirmed from the literature that there was no particular change in protonation state of His residues associated with fragment binding. The binding sites for docking were defined for each receptor as an 8.0Å sphere from the centre of the fragment binding positions.

4.4.2 Preparation of solvent probes and fragments

The bound solvent and fragment molecules were extracted from the original PDB files and stored as separate SD files. The coordinates were extracted from the original PDB files and visually inspected to correct bond orders and hydrogen atoms added to complete valency. The atom types in the fragment were assigned based on connectivity and bond order and parameters assigned from the CHARMm Momany and Rone forcefield[182] and assigned MMFF94 charges. The resulting atom types for

fragments and partial charges are shown in Appendix (Section 7.1). Missing forcefield parameters were estimated automatically based on similar combinations already available in the parameter list. Hydrogen atom positions were then optimized using CHARMm[183] and the fragment structures minimized with ABNR minimization to a gradient of 0.1 kcal $mol^{-1}Å^{-1}$. The minimized structures were then used as input for the docking procedures.

4.4.3 MCSS Minimization

In a typical MCSS protocol, several copies (up to 1000) of a functional group are placed in the binding site and then simultaneously energy-minimized such that copies experience only the field from the protein. If more than one copy converges to the same position within a specified distance threshold, only one copy is retained. At the end, a collection of energy minima is obtained, each of which is associated with a position, interaction geometry and energy score. In this study, 750 copies of each fragment were placed inside the binding site of corresponding receptor structure. A distance-dependent dielectric model was used to approximate the solvent, using a dielectric constant of 1.

The energy minimization was carried out by performing an initial 500 steps of steepest descent followed by 300 steps of additional steepest descent and then 20 repetitions of 500 steps of conjugate gradient minimization. At each repetition, a single copy was retained where multiple copies converged within an RMSD of 0.2Å. At the end of the MCSS run, energy minima within 2.0Å RMSD of each other were considered as one cluster and the minimum with the highest score were selected as the cluster representative.

For the fragment docking dataset, MCSS calculations were performed for each fragment in its own receptor. For the HSP90 dataset set cross-docking calculations were also performed by running MCSS calculations for each fragment against all receptors in the binding site sphere defined by the position of the native fragment.

4.4.4 Minimization of fragment poses

As a post-processing step, MCSS and GOLD poses were minimized in the context of the target binding site with ABNR minimization to a gradient of 0.1 kcal $mol^{-1}Å^{-1}$. The protein was held rigid while the fragments were fully flexible. The native position of the fragment was also minimized under the same conditions to give the *in situ* minimised X-ray pose.

4.4.5 Docking with GOLD

For the fragment docking dataset, docking and scoring was also performed with GOLD[184] by Hugues-Olivier Bertrand of Accelrys. The binding site sphere defined previously for each receptor structure was used and 40 docking runs were performed for each fragment. Default values were used for genetic algorithm parameters, "Generate Diverse Solutions" was set to TRUE (Cluster Size = 2, RMSD =1) and solutions were scored using the GOLDScore fitness function.

4.4.6 MM/GBMV-SA Scoring Scheme

The binding free energy of energy minima obtained from MCSS was evaluated using a variation of the standard MM-GB/SA approach[47]. The outline of this scoring scheme is shown in Figure 4.4.

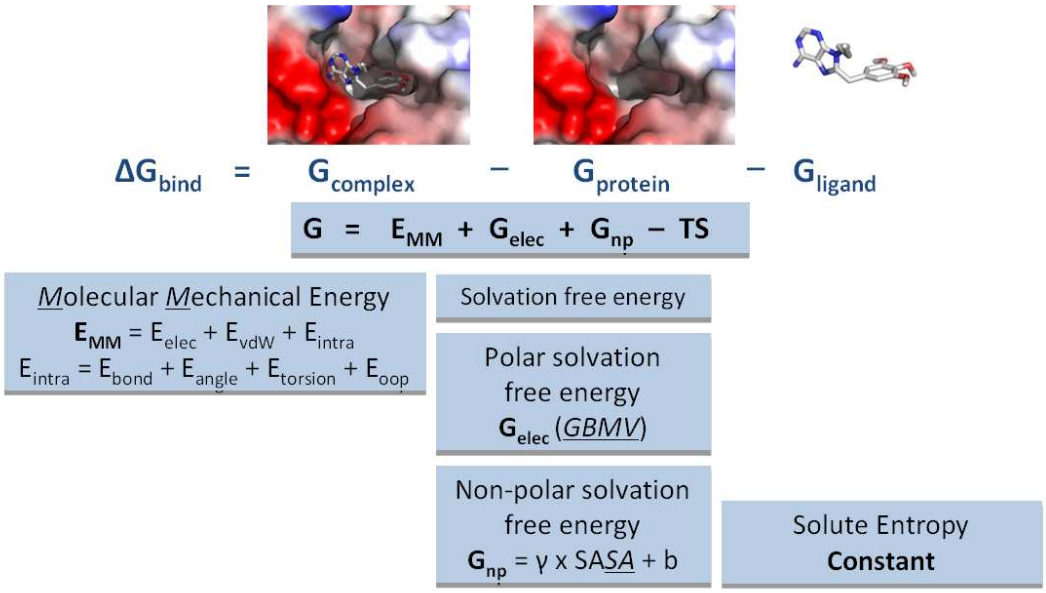


Figure 4.4. Summary of the scoring scheme (MM/GBMV-SA) used in this study.

For each pose obtained from MCSS or GOLD, the binding free energy is evaluated as:

$$\Delta G_{bind} = G_{complex} - G_{protein} - G_{ligand}$$

The free energy of each of the above terms is calculated from:

$$G = E_{MM} + G_{elec} + G_{np} - TS$$

E_{MM} is the molecular mechanical energy calculated from CHARMM force-field, G_{elec} and G_{np} represent electrostatic and non polar components of solvation free energy. TS represents the solute entropy which, in this study, was assumed to be constant among a set of poses for the same ligand in a binding site. The mathematical expressions for each of these terms are presented in Table 4.7. E_{MM} represents the gas-phase forcefield energy and consists of internal energy (E_{int}), electrostatic energy (E_{elec}) and van der Waals energy components. E_{int} is further divided into E_{bond}, E_{angle}, $E_{torsion}$ and E_{oop} to take into account energies associated with bonds, angles, torsions and out of plane motions. The electrostatic component, E_{elec} is calculated from

Coulomb's expression using a dielectric constant of 1 and van der Waals energy, E_{vdW}, is calculated from the Lennard-Jones 6-12 potential.

The electrostatic component of the solvation free energy (G_{elec}) was calculated using the Generalized Born method with Molecular Volume integration (GBMV)[50, 101, 107]. GB methods are semi-analytical approximations to the more rigorous PB equation and they have been shown to reproduce electrostatic solvation energies obtained from the latter, with an error rate of ≤ 1%[107]. The expression used for the electrostatic solvation energy under GB formalisms is:

$$\Delta G_{elec} = -\frac{1}{2}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right)\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{i,j} + \alpha_i \alpha_j \exp(-r_{i,j}^2 / F\alpha_i \alpha_j)}}$$

where ε_p and ε_w represent solute and solvent dielectric constants, $r_{i,j}$ is the distance between atoms i and j, α_i is the GB radius of atom i. The factor F is a scaling factor for GB radii, whose most commonly used value is 4. The GBMV approach uses a numerical integration of molecular volume to calculate Born radii[100, 107]. In this study, for the solute and the solvent, dielectric constants of 1 and 80 were used, respectively. The non-polar contribution (G_{np}) to solvation free energy was calculated based on Surface Area (SA) model which assumes a linear relationship between G_{np} and the solvent accessible surface area. The values for constants, γ and β, were set to 0.00542 Kcal/molÅ and 0.92 Kcal/mol, respectively.

Table 4.7. Energy terms in MM/GBMV scoring scheme used in this study.

Term	Mathematical Expression	Notes
E_{MM}	$E_{bond} = \sum k_b (r - r_0)^2$	Bond potential
	$E_{angle} = \sum k_\theta (\theta - \theta_0)^2$	Angle potential
	$E_{tor} = \sum \lvert k_\varphi \rvert - k_\varphi \cos(n\varphi)$	Torsional potential
	$E_{oop} = \sum k_\omega (\omega - \omega_0)^2$	Out-of-plane motion potential
	$E_{vdw} = \sum_{i>j=1} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) sw(r_{ij}^2, r_{on}^2, r_{off}^2)$	Van der Waals potential: based on Lennard-Jones 6-12 potential, sw() is a switching function to control the size of non-bonded list
	$E_{elec} = \sum_{i>j=1} \frac{q_i q_j}{\varepsilon r_{ij}}$	Electrostatic potential: based on Columbic interactions, In MCSS a distance-dependent model is used where ε is dielectric constant and r is the distance between two atoms
G_{elec}	$\Delta G_{elec} = -\frac{1}{2}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w} \right) \sum_{i,j} \frac{q_i q_j}{\sqrt{r_{i,j} + \alpha_i \alpha_j \exp(-r_{i,j}^2 / F\alpha_i \alpha_j)}}$	Electrostatic component of free energy of solvation: standard form of the GB model, details are described in 4.1.4
G_{np}	$\Delta G_{np} = \gamma SASA + b$	Non-polar component of free energy of solvation

4.5. Results

The output of an MCSS calculation is a set of poses for each probe or fragment molecules inside protein binding sites. Each of these poses has an MCSS score - a binding energy that is calculated based on the CHARMm forcefield (Table 4.7). These poses are ranked according to their MCSS scores, with the highest scoring pose representing the most energetically favourable one. This pose is actually a representative of a cluster that results from the convergence of multiple functional group copies to a similar position in the binding site. The top-scoring pose within the cluster is chosen as the representative.

When the experimentally determined position of a ligand is known then the validation of MCSS calculations is performed by comparing the root mean square deviation (RMSD) of predicted pose against the reference (experimental) pose. The

experimental positions of ligands are derived from X-ray model building and refinement. An energy minimization protocol applied to the X-ray poses inside the binding site could result in a deviation from the original position, which could be a result of either force-field limitations or poor ligand placement in the X-ray structure. Therefore, in this study, the *in situ* minimized pose was also considered as the reference.

This chapter presents the results of MCSS calculations on solvent mapping datasets (elastase and thermolysin) and their comparison with another solvent mapping algorithm (FT-Map).

4.6 MCSS calculations on Elastase Dataset

The results of MCSS calculations on elastase are summarized in Table 4.8. For each solvent position, the RMSD from the X-ray position ($RMSD_{X\text{-}ray}$) of the nearest predicted pose along with its rank and MCSS score is listed. Out of 14 solvent positions in the active site cleft, nine were predicted at an $RMSD_{X\text{-}ray}$ equal to or less than 2.0Å. None of the predicted poses that were closest to the experimental binding position was given the highest rank.

The experimental solvent mapping studies indicate that almost all of the solvent probes bound to the S1 sub-site and such clustering of solvent probes could reflect ligand binding sites[82]. The predictions for this site are, therefore, particularly important. It was noticed that for five solvent positions in S1, three were predicted with $RMSD_{X\text{-}ray} \leq 2.0$Å. The S4 sub-site is lined with hydrophobic residues and has a preference for apolar residues in the substrate[185]. It was noticed that one out of three solvent positions in S4 sub-site were predicted with $RMSD_{X\text{-}ray} \leq 2.0$Å. Similarly,

for S1' and S3' sub-sites which also provide hydrophobic contacts to bound substrates[82], two out of four positions were predicted within the $RMSD_{X\text{-}ray}$ cut-off. Finally, for the oxyanion hole, represented by the S3 sub-site and characterized by mainly polar interactions[82], all three experimental solvent positions were predicted at $RMSD_{X\text{-}ray} \leq 2.0$Å.

Table 4.8. Results of MCSS calculations on Elastase for different solvent probes in their native protein structures.

Solvent	PDB	Sub- site	Nearest Pose		
			$RMSD_{X\text{-}ray}$	Rank	Score
ACN1001	2FO9	S1	2.19	31	10.87
IPA1001	2FOA	S1	3.21	27	13.45
IPA1002	2FOA	S4	2.29	52	10.36
IPA1003	2FOA	S3	1.93	3	17.86
DMF1004	2FOC	S3'	0.66	11	19.96
EOH1001	2FOD	S1	1.88	11	15.17
EOH1002	2FOD	S4	1.87	22	13.18
EOH1003	2FOD	S3	1.88	13	15.06
EOH1004	2FOD	S3'	1.85	51	9.99
HEX1001	2FOE	S1	1.31	54	7.12
HEX1004	2FOE	S3'	3.06	43	8.25
TFE1001	2FOG	S1	2.07	4	15.13
TFE1002	2FOG	S4	2.02	30	9.74
TFE1003	2FOG	S3	1.56	6	13.57
TFE1008	2FOG	S1'	3.47	32	9.69

MCSS calculations were also performed with the same solvent probes but on a generic thermolysin structure in order to investigate the effect of starting conformation of the receptor. The structure (2FO9) was chosen as the one which had the minimum average pair-wise backbone RMSD with all other structures in the dataset. The results are summarized in Table 4.9.

Table 4.9. Results of MCSS calculations on Elastase for solvents in a generic (2FO9) receptor structure.

Solvent	Sub- site	Nearest Pose		
		$RMSD_{X\text{-}ray}$	Rank	Score
ACN1001	S1	2.19	31	10.87
IPA1001	S1	0.52	12	16.81
IPA1002	S4	2.12	19	14.71
IPA1003	S3	1.82	3	14.62
DMF1004	S3’	0.87	9	19.72
EOH1001	S1	1.98	14	15.00
EOH1002	S4	1.58	21	13.14
EOH1003	S3	1.44	11	15.21
EOH1004	S3’	1.96	46	10.10
HEX1001	S1	3.08	68	5.03
HEX1004	S3’	8.81	61	6.11
TFE1001	S1	1.63	15	11.35
TFE1002	S4	2.71	19	10.90
TFE1003	S3	0.94	2	13.17
TFE1008	S1’	2.38	26	9.78

Although the number of solvent positions predicted within the $RMSD_{X\text{-}ray} \leq 2.0$Å are similar, the composition of these predictions is slightly different (Table 4.9). The nearest predicted pose for TFE1003 has a higher $RMSD_{X\text{-}ray}$ in the generic receptor than in native receptor. On the other hand, the nearest predicted pose 1PA1001 gets a better rank and has much lower $RMSD_{X\text{-}ray}$ in the generic receptor than in the native receptor. Apart from these cases, the overall results were similar to those obtained for native receptors (Table 4.8).

The output of MCSS was also analysed with the *in situ* minimized X-ray pose as reference. It should be noted that in some cases *in situ* minimization resulted in significant deviation from the X-ray positions of solvents, as indicated by $RMSD_{X\text{-}ray|X\text{-}rayMin}$ column in Table 4.10.

Table 4.10. Results of MCSS calculations on Elastase using *in situ* minimized poses as reference. For each solvent position, the RMSD of *in situ* minimized pose from the X-ray pose ($RMSD_{X\text{-}ray|X\text{-}rayMin}$) is given whereas the nearest predicted pose with respect to *in situ* minimized pose ($RMSD_{X\text{-}rayMin}$) is shown with its ranks and score.

Solvent	PDB	$RMSD_{X\text{-}ray\|X\text{-}rayMin}$	$RMSD_{X\text{-}rayMin}$	Rank	Score
ACN1001	2FO9	0.73	2.04	31	10.87
IPA1001	2FOA	0.53	3.60	27	13.45
IPA1002	2FOA	2.04	1.88	52	10.36
IPA1003	2FOA	0.98	1.42	3	17.85
DMF1004	2FOC	0.77	0.67	11	19.96
EOH1001	2FOD	0.84	2.03	44	10.79
EOH1002	2FOD	1.88	0.06	22	13.18
EOH1003	2FOD	1.50	1.77	13	15.06
EOH1004	2FOD	1.01	1.45	51	9.99
HEX1001	2FOE	0.71	1.24	54	7.12
HEX1004	2FOE	0.60	3.14	43	8.25
TFE1001	2FOG	0.55	2.23	4	15.13
TFE1002	2FOG	1.57	0.87	30	9.74
TFE1003	2FOG	1.37	1.29	6	13.57
TFE1008	2FOG	1.56	3.22	50	6.78

For example the *in situ* minimized pose of IPA1002 in 2FOA moves deeper into the S4 sub-site where it can make hydrogen bonding interactions with Arg226 and Val224 (Figure 4.5) and probably better hydrophobic contacts with Phe223 and Ala104. In such a case, the output of MCSS is expected to be closer to the X-ray *in situ* minimized pose, showing lower RMSD of the nearest cluster from *in situ* minimized pose than RMSD from the X-ray pose. This was noticed from the comparison of $RMSD_{X\text{-}ray}$ and $RMSD_{X\text{-}rayMin}$ for IPA1002, EOH1002, EOH1003, EOH1004, TFE1002, TFE003, TFE1008 in Table 4.8 and 4.10.

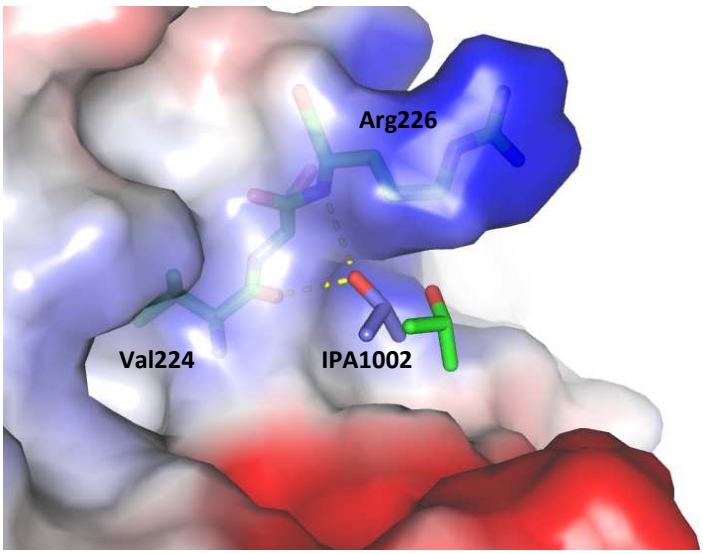


Figure 4.5. Change in the X-ray position of IPA1002 (green C-atoms) after *in situ* minimization (blue C-atoms).

4.7 Comparison with experimental positions

4.7.1 Acetone

Organic solvents were shown to bind to at least 6 sites in the elastase active site (S1, S2, S3, S4, S1' and S3')[82]. In the active site acetone bound only to S1 but three other sites were also observed, two of which were crystal contacts and one was a potential interaction site (named ACN1006). Although this site is located quite far from the active site cleft, it is still inside the binding site sphere used in this study. The top scoring pose predicted from MCSS had $RMSD_{X\text{-}ray}$ of 2.8Å with respect to this position and reproduced the same interactions as observed for ACN1006 experimentally[82], hydrogen bonding with the Lys234 Nε atom and hydrophobic interactions with Leu227 (Figure 4.6). In the active site, the nearest predicted pose for ACN1001 was ranked 31 with an $RMSD_{X\text{-}ray}$ 2.19Å. A hydrogen bond with Ser225 side-chain was observable which was also noted in the X-ray structure[82].

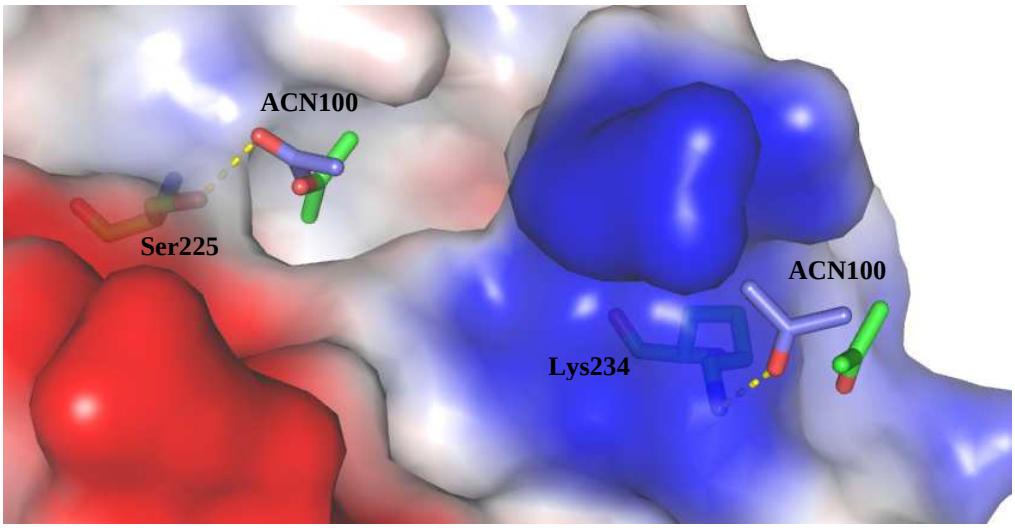


Figure 4.6. X-ray (green C-atoms) and nearest MCSS poses (blue C-atoms) generated for ACN1001 and ACN1006.

4.7.2 Iso-propanol

IPA binds to three positions in binding sites S1 (IPA1001), S4 (IPA1002) and S3 (IPA1003). The top-scoring pose from MCSS was located in a region which is occupied by water molecules in the X-ray structure, making similar hydrogen bonding interactions to protein side-chains (Figure 4.7).

It was noticed that, in general, the solvent positions in the oxyanion hole (the S3 sub-site) were reproduced relatively accurately as compared to other sub-sites (Table 4.8).

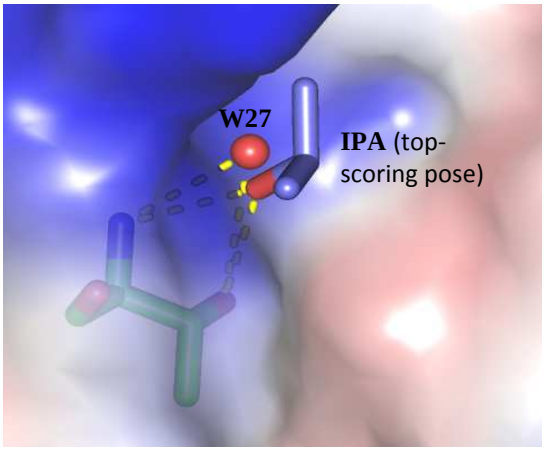


Figure 4.7. Top-scoring IPA pose generated by MCSS and a corresponding bound water molecule in the X-ray structure.

For IPA1001, the nearest predicted pose ($RMSD_{X\text{-}ray}$ 3.21Å) (Figure 4.8B) was ranked 27 in the native structure. Surprisingly, in the generic receptor structure, the X-ray pose was reproduced at very low $RMSD_{X\text{-}ray}$, 0.52Å (Table 4.9). The pair-wise RMSD between these two receptor structures is only 0.26Å.

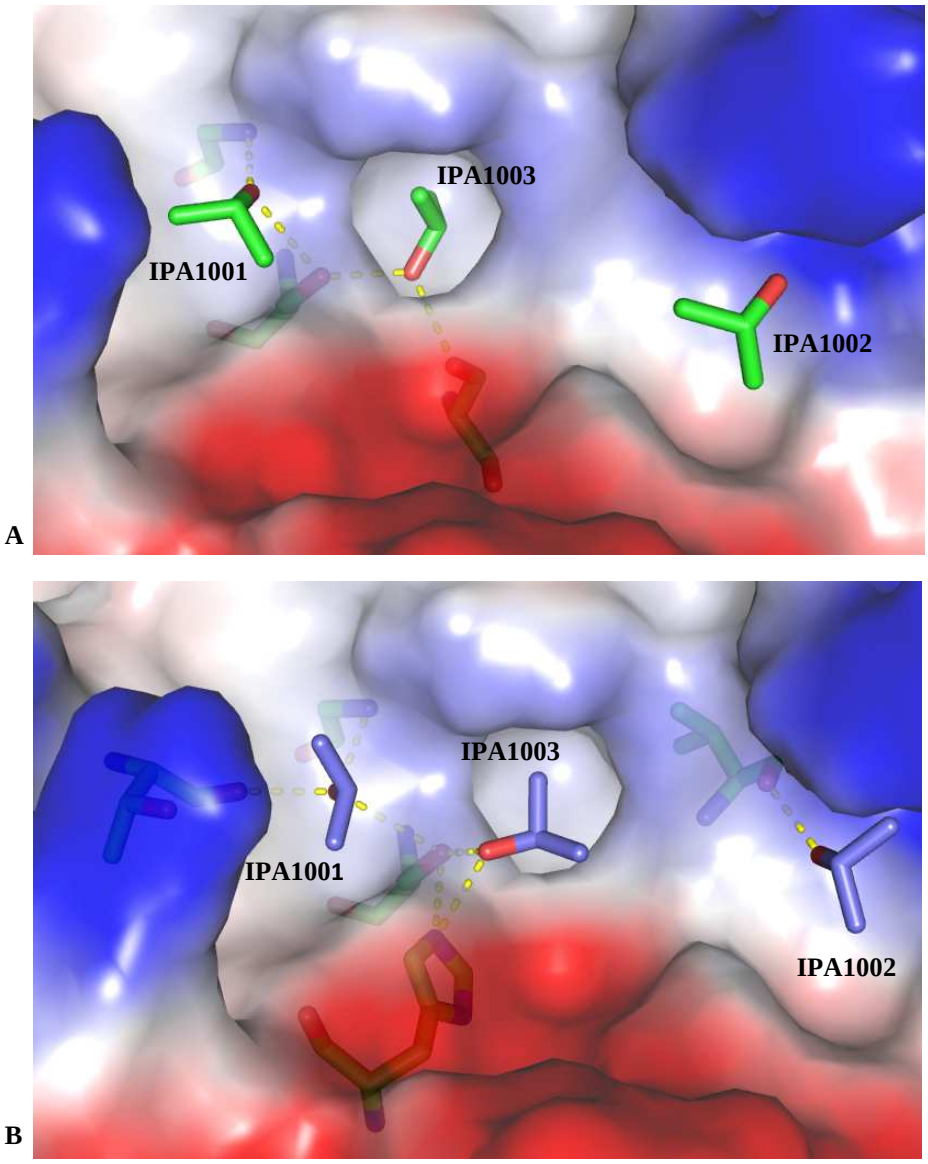


Figure 4.8. MCSS predicted poses for IPA. A. X-ray (green C-atoms) and B. nearest MCSS (blue C-atoms) poses generated for IPA1001, IPA1002 and IPA1003.

IPA1002 represents a special case as it is the solvent position with the highest deviation from the X-ray pose during *in situ* minimization (Table 4.10). In the original structure, this pose is suggested to make hydrophobic contacts in S1 sub-site

involving Phe223, Val102, Ala104 and Thr182[82] and possibly hydrogen bond with Val224 main-chain carbonyl group (Figure 4.8A). As discussed before, it was observed that *in situ* minimization resulted in a pose that makes hydrogen bond with good geometry and buried slightly deeper into the hydrophobic patches in S1 than the X-ray pose (Figure 4.5).

The average atomic B-factor of IPA1001 (47.15Å^2) is also slightly higher than the average atomic B-factors of all other solvent molecules bound at S1 (39.38 Å^2). If the minimized pose is considered as reference then a relatively closer pose is predicted ($RMSD_{X\text{-}ray\ in}$ 1.81Å) (Table 4.10) at rank 52.

The third experimentally observed position of IPA (IPA1003) lies in the oxyanion hole (the S3 sub-site) and it was observed that as for other solvent positions in oxyanion hole, IPA1003 was also reproduced relatively accurately with 1.93Å $RMSD_{X\text{-}ray}$ and at rank 3 (Table 4.8) (Figure 4.8A and B).

4.7.3 Dimethylformamide (DMF)

DMF binds only at S3' sub-site in the active site[82]. MCSS calculations resulted in a DMF pose with $RMSD_{X\text{-}ray}$ value of 0.66 and rank 11. The top-scoring pose did not represent any other experimentally relevant positions or water binding sites.

4.7.4 Ethanol (EOH)

Ethanol is shown to bind to 4 sub-sites in the active site cleft[82], S1, S3, S4 and S3'. MCSS results indicate that for all these interactions sites solvent binding modes were reproduced at $RMSD_{X\text{-}ray} \leq 2.0$Å (Table 4.8). However, the binding orientations of predicted poses are slightly different in most of the cases. A very noticeable feature of the poses predicted nearest to all EOH sites is that they are more extensively

hydrogen bonded to protein side-chains than their X-ray counterparts (Figure 4.9A and B).

For EOH1001, the X-ray pose is oriented in a way that points non-polar –CH3 group into the hydrophobic pocket in S1 and hydroxyl group is exposed to the solvent. Due to the absence of detailed treatment of water mediated interactions in the MCSS scoring scheme, the predicted pose at this position favours a binding mode where the hydroxyl group forms a hydrogen bond with the Ser203 side-chain (Figure 4.9A and B). This contributes mainly to the observed $RMSD_{X\text{-}ray}$ as the distance between the methyl carbon of X-ray and predicted poses (1.9Å) is less than the distance between hydroxyl oxygen (2.4Å) of the two poses.

Similarly, at S4, the predicted and X-ray pose of EOH1002, shows slightly different binding orientations. The predicted pose shows favourable hydrogen bonding with backbone groups of Val224 and Arg226, which is absent in the case of X-ray pose (Figure 4.9A and B). In solution, such interactions are not always favourable because of the screening effect of high dielectric solvent[81]. For EOH1003 and EOH1004, the binding orientation of predicted poses is also favoured towards hydrogen bonding interactions, causing deviation in the hydrophobic contacts (Figure 4.9A and B). This is further strengthened by the observed deviation of *in situ* minimized poses from X-ray poses for EOH, as indicated in Table 4.10. It should be noted that the *in situ* minimization protocol also uses a distance-dependent dielectric model for solvent approximation therefore similarity in the outputs generated by *in situ* minimization protocol and MCSS is expected.

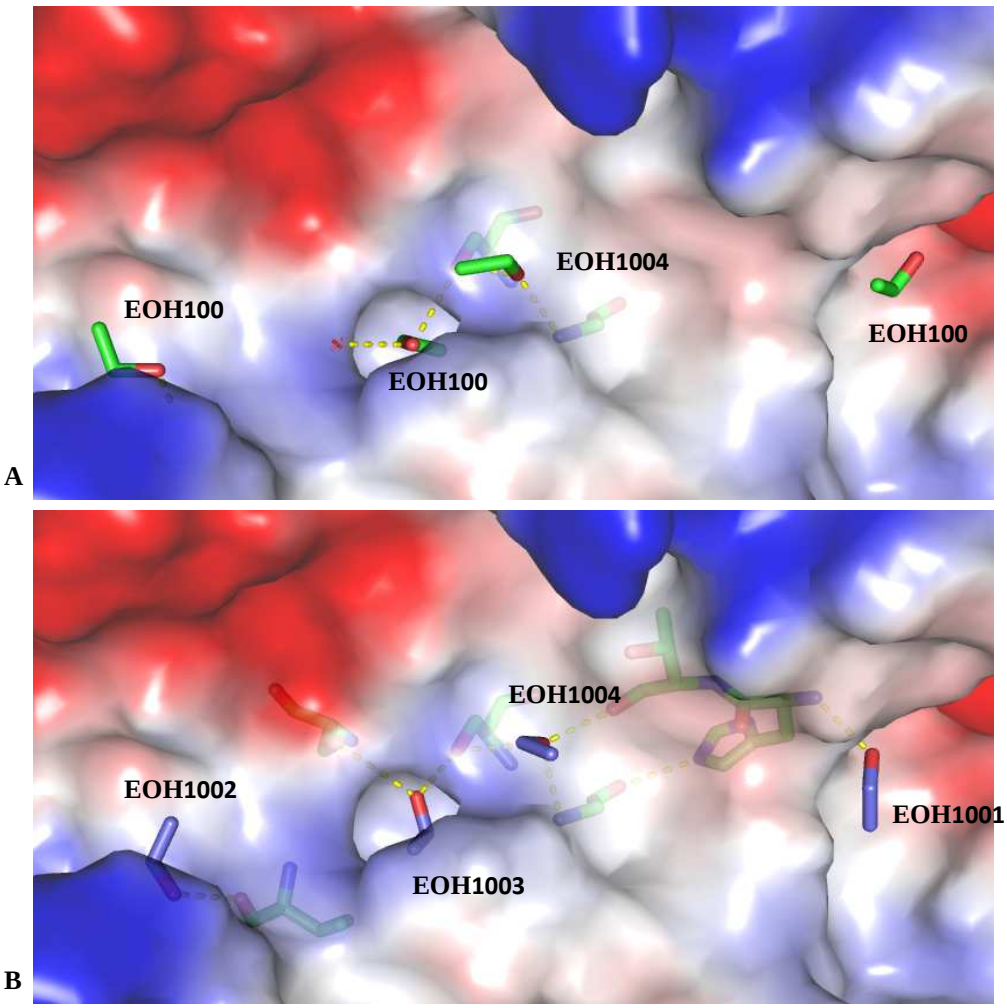


Figure 4.9. MCSS calculations with EOH probe. A. X-ray poses (green C-atoms) and B. nearest MCSS poses (blue C-atoms) generated for EOH1001, EOH1002 and EOH1003 and EOH1004.

4.7.5 5-Hexene-1,2-diol (HEX)

HEX is the largest solvent probe in the dataset and consists of two hydroxyl groups and a relatively large non-polar part, as compared to other solvent probes. The experimental binding sites of HEX include S1 and S3' which provide hydrophobic pockets for alkene side-chain and hydroxyl groups are projected to the solvent[82]. For HEX1001, this binding mode was predicted by MCSS at RMSDX-ray 1.31Å, at rank 54. Among the top-scoring poses of HEX, at least two were located in S3 where no

experimental pose of HEX was observed. In the original structure solved in 80% HEX[82], there is a sulphate ion bound at this site making similar interaction as observed for top-scoring pose (Figure 4.10). It was previously noted that the binding at S3 sub-site is driven mainly by polar interactions. The most favourable pose of HEX was predicted at this site.

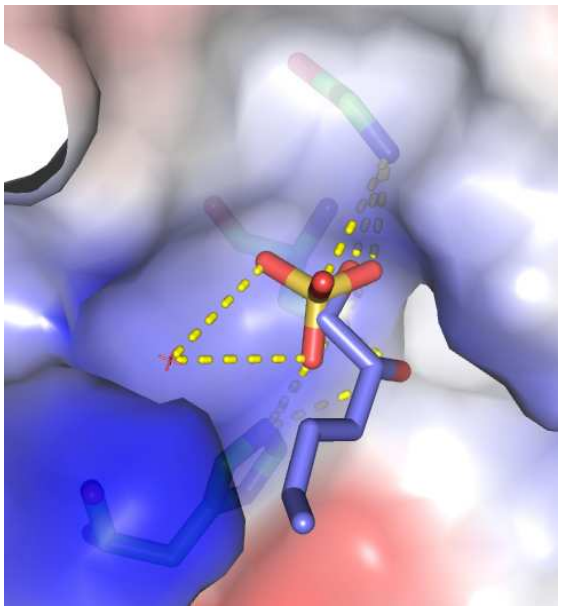


Figure 4.10. The top-scoring pose for HEX at S3 sub-site. In the original structure a sulphate ion was bound at this site.

For HEX1004, a binding mode was predicted close to the X-ray position but with $RMSD_{X\text{-}ray}$ 3.47Å. The non-polar side-chain was oriented in a completely different manner, missing key hydrophobic contacts with Leu77 which were observed in the X-ray pose.

4.7.6 Trifluoroethanol (TFE)

The experimental binding sites for TFE are located in S1 (TFE1001), S3 (TFE1003), S4 (TFE1002) and S1' (TFE1008) sub-sites. The clusters 4 and 6 from MCSS correspond to TFE1001 and TFE1003, respectively. The observed deviation in the predicted

binding modes mainly results from the different orientation of the hydroxyl group from the X-ray pose (Figure 4.11A and B). This is consistent with observations for similar solvent probes noted previously and is a consequence of the force-field favouring the hydrogen bonding for the hydroxyl within the S1 sub-site, instead of projecting it towards the solvent.

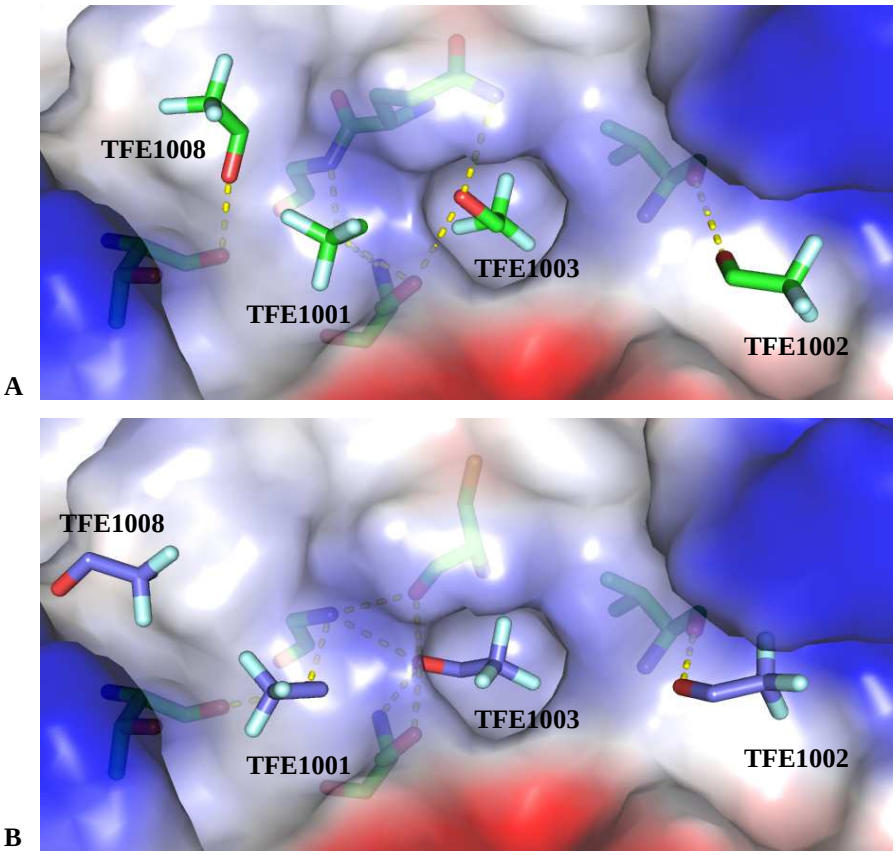


Figure 4.11. MCSS calculations with TFE probe. A. X-ray poses (green C-atoms) and B. nearest MCSS poses (blue C-atoms) generated for all crystallographically observed poses are shown.

For TFE1002, on the other hand, the nearest predicted pose ($RMSD_{X\text{-}ray}$ 2.02Å) reproduced key interactions as observed in the X-ray pose and was ranked 30. For TFE1008, the nearest predicted pose was still far away from the X-ray pose ($RMSD_{X\text{-}}$

$_{ray}$ 3.47Å) (Figure 4.11). One of the fluorine atoms in this predicted binding mode was located at the position of a water molecule in the original structure. The absence of waters in the calculations could possibly be the reason for MCSS favouring the binding mode at this position deviated towards a nearby water molecule.

4.8 MCSS calculations on Thermolysin

The results of MCSS calculations on thermolysin are summarized in Table 4.11. For each solvent position, the RMSD from the X-ray position ($RMSD_{X\text{-}ray}$) of the nearest predicted pose along with its rank and MCSS score is indicated. It was observed that out of seven solvent molecules in the active site cleft, the binding modes of four molecules were predicted at $RMSD_{X\text{-}ray}$ equal to or less than 2.0Å. Three of these predicted poses were also ranked 1 or 2 among the candidate poses.

Table 4.11. Results of MCSS calculations on Thermolysin for solvents in their native protein structures. The predicted poses with the lowest RMSD from X-ray ($RMSD_{X\text{-}ray}$) are shown with their ranks and scores.

Solvent	PDB	Sub- site	Nearest Pose		
			$RMSD_{X\text{-}ray}$	Rank	Score
ACN1	1FJQ	S1’	1.72	1	26.92
CCN1	1FJU	S1’	2.26	1	23.18
IPH1	1FJW	S1’	0.61	2	22.39
IPA1	8TLI	S1’	1.61	10	15.05
IPA5	8TLI	S8	2.19	17	13.41
IPA8	8TLI	S5	1.53	11	14.77
IPA9	8TLI	S2’	2.31	20	12.58

The experimental studies performed by English *et al.* (2001) showed that almost all of the solvent probes bound to the S1’ sub-site, the main specificity pocket for the enzyme. The predictions for this site are, therefore, particularly important. It was noticed that most of the correct predictions corresponded to solvent positions in S1’. Only for one probe (CCN), the binding mode at S1’ was predicted with relatively

higher $RMSD_{X\text{-}ray}$ (Table 4.11). Additional solvent binding sites were considered for IPA, two of which, IPA1 and IPA8, were predicted within $RMSD_{X\text{-}ray}$ cut-off of 2.0Å.

In order to test the sensitivity of the method towards the starting conformation of the receptor, the same calculations were performed on a generic thermolysin structure (PDB code: 2TLX) and results are shown in Table 4.12. In general, the results are somewhat similar in terms of the number of solvent positions reproduced with reasonable $RMSD_{X\text{-}ray}$. In some case, slight differences can be observed both in ranking and the $RMSD_{X\text{-}ray}$ of the nearest pose. For instance, the nearest predicted pose for IPH1 in the generic receptor structure had lower $RMSD_{X\text{-}ray}$ and poorer rank then in the native structure. On the other hand, the nearest predicted pose 1PA1 had a better rank and a much lower $RMSD_{X\text{-}ray}$ in the generic receptor than in the native receptor (Table 4.12).

Table 4.12. Results of MCSS calculations on Thermolysin for solvents in generic (2TLX) solvent-bound structure.

Solvent	Sub-site	Nearest Pose		
		$RMSD_{X\text{-}ray}$	Rank	Score
ACN1	S1’	0.74	1	26.13
CCN1	S1’	2.20	1	22.19
IPH1	S1’	2.24	3	23.20
IPA1	S1’	1.36	2	22.99
IPA5	S8	2.43	11	14.36
IPA8	S5	1.68	9	16.41
IPA9	S2’	2.20	21	11.76

The *in situ* minimization of X-ray poses of solvent resulted in significant deviations for four cases, as indicated by $RMSD_{X\text{-}ray|X\text{-}rayMin}$ values in Table 4.13. When this is compared with the results in Table 4.11, it appears that, in all such cases using X-ray pose does not give correct predictions. Consequently, a comparison of the predicted poses with *in situ* minimized poses shows better RMSD values and ranks (Table 4.13).

Table 4.13. Results of MCSS calculations on Thermolysin using *in situ* minimized poses as reference.

Solvent	PDB	$RMSD_{X\text{-}ray|X\text{-}rayMin}$	$RMSD_{X\text{-}rayMin}$	Rank	Score
ACN1	1FJQ	1.20	0.99	1	26.099
CCN1	1FJU	2.25	0.87	1	23.181
IPH1	1FJW	0.56	2.47	2	19.942
IPA1	8TLI	2.02	1.11	9	15.054
IPA5	8TLI	1.04	2.02	4	16.569
IPA8	8TLI	0.71	0.98	7	15.648
IPA9	8TLI	1.75	2.55	2	17.606

4.9 Comparison with experimental positions

All the solvent probes used in the experimental solvent mapping studies for thermolysin[81] have a hydrophilic and a hydrophobic part. The S1' sub-site contains hydrophobic residues at the base of the pocket (Phe130, Leu133, Val139, Ile188, Val192 and Leu202) and polar residues (Asn112, Glu143, Arg203 and His231) towards the edges. The experimentally determined binding mode for all the solvent probes at S1' sub-site is roughly similar in terms of the orientation of polar and non-polar groups. The poses predicted from MCSS at this sub-site reproduce similar binding geometries for ACN1, IPA1 and IPH1 (Figure 4.12). In the case of CCN1, although a high-scoring pose is predicted at the same location but with almost opposite orientations of methyl and nitrile groups. It was notable that in the original X-ray structure, the CCN1 has a high atomic B-factor and was expected to be mobile[81]. Furthermore, the *in situ* minimization also resulted in a pose with large deviation from the X-ray pose (Table 4.13).

IPA also binds to three additional sub-sites in thermolysin binding site, denoted by IPA5, IPA8 and IPA9 (Figure 4.12). These interactions were observed at high concentration (90% isopropanol) (English *et al.* 2001) and interestingly the quality of

prediction from MCSS for these sites was reduced compared to IPA1, as seen in $RMSD_{X\text{-}ray}$ values and ranks (Table 4.11). It was noted the observed $RMSD_{X\text{-}ray}$ of the nearest predicted poses for IPA5 and IPA9 results from MCSS favouring a binding mode with distinct hydrogen bonding interactions.

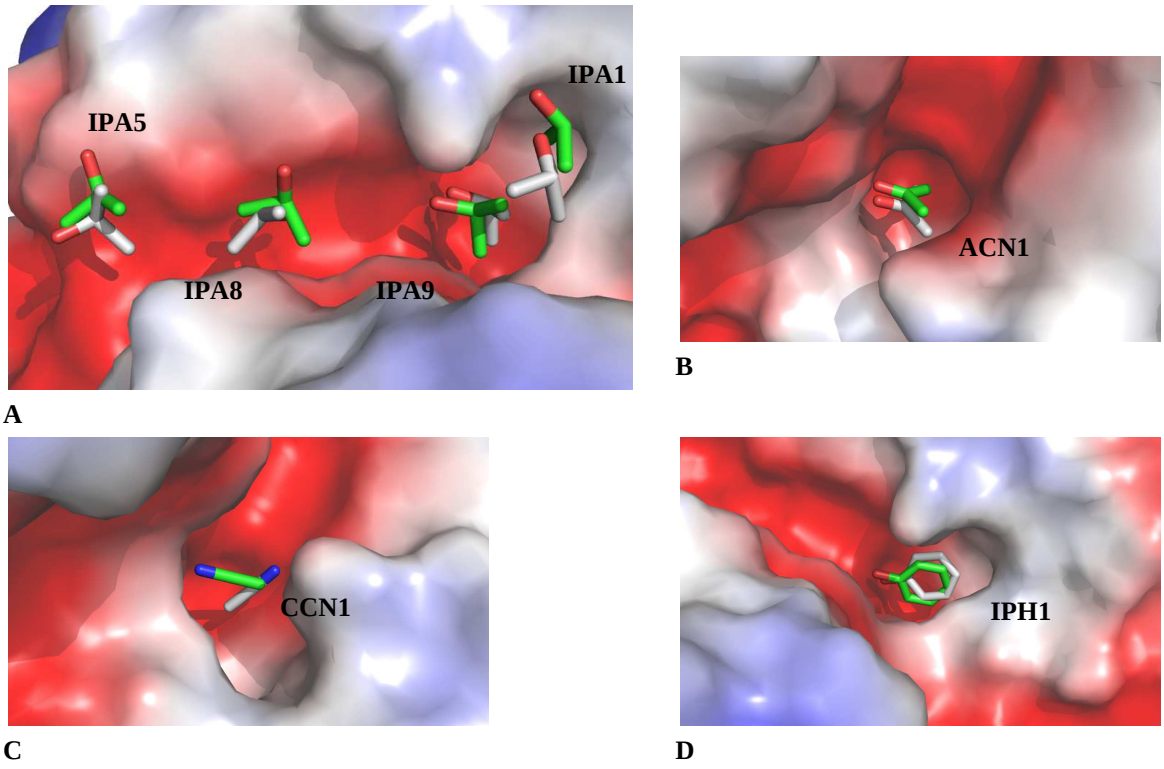


Figure 4.12. The X-ray (green C-atoms) and predicted poses (grey C-atoms) for solvent probes in thermolysin binding site. For each solvent the nearest pose predicted from MCSS is shown. A. IPA (isopropanol), B. ACN (acetone), C. CCN (acetonitrile), D. IPH (phenol)

4.10 Discussion

MCSS was developed primarily as a technique to probe binding sites for energetically favourable positions of functional groups. As discussed above, the experimental solvent mapping data is a very useful resource in evaluating MCSS predictions. Here we also discuss a comparison of MCSS calculations with another solvent mapping

algorithm. The results of computational solvent mapping for elastase and thermolysin based on FT-Map were reported by Brenke *et al.*[88]. A brief overview of FT-Map method was described earlier (Section 1.4.3) and the most important difference between the MCSS and FT-Map lies in the way electrostatic interactions are taken into account by these two methods. FT-Map evaluates electrostatic and solvation terms for solvent binding modes using Analytic Continuum Electrostatic model implemented in CHARMM 27 version[88] whereas MCSS energy minimization uses a distance-dependent dielectric model. A set of 16 solvent probes was applied on an inhibitor-bound structure of elastase (PDB code: 2ELA) and a dipeptide-bound structure of thermolysin (PDB code: 2TLX). For each solvent probe, the six lowest free energy clusters were superimposed and compared with experimental probe positions to delineate consensus sites, which have been experimentally shown to correspond to substrate binding pockets of proteins[78]. For elastase, the largest consensus site, which contained 20 clusters represented by all 16 probes, was identified in the S1 sub-site. Similarly, for the other four sub-sites (S3, S4, S1' and S3') clusters represented by four up to nine different probe molecules were obtained. For thermolysin, the largest consensus site contained 19 clusters, represented by all 16 probes and the centre of this site, the lowest energy cluster, corresponded to the main specificity pocket, S1'. Three additional sites, adjacent to S1', contained 16, 7 and 8, probe clusters, respectively and trace out S1 and S2 sites collectively (Brenke *et al.* 2009).

These results are in qualitative agreement with the results from MCSS in this study. For elastase, out of 5 solvent probes, MCSS reproduced clusters represented by 2 to 3 different probes in each of the solvent binding sub-sites. Similarly, for thermolysin,

high scoring clusters for all four probes were predicted in S1'. The main difference between the two methods is reflected in the ranking of clusters from MCSS. The probe clusters in the case of FT-Map that line the binding site are always within the six lowest free energy clusters whereas in the case of MCSS the scoring and ranking are not so efficient, particularly in the case of elastase. For thermolysin, the ranking of probe clusters nearest to experimental solvent binding sites is slightly better. English *et al.* compared experimental results with MCSS results for thermolysin and highlighted the absence of desolvation terms as a possible reason for the appearance of false positive energy minima and poor ranking[81]. Additionally, Silberstein *et al.* using FT-Map, showed that ranking of poses could be improved by taking the mean score for each cluster and choosing the pose closest to the mean value as cluster representative. MCSS on the other hand takes the highest scoring pose within a cluster as the cluster representative. We therefore repeated the calculations, described in 5.2 and 5.4, with a different ranking strategy where at the end of MCSS, the mean value pose was chosen as the cluster representative. The results of these calculations are shown in Appendix (Table 7.1 and 7.2).

It was observed that ranking improved in some cases, however, in other cases it remained unchanged or even deteriorated. The overall effect of re-ranking therefore did not significantly improve the quality of predictions. In principle at a low clustering RMSD, the difference among poses within a cluster should not be significant. In previous application of MCSS on thermolysin[81], a clustering RMSD of 3.5Å was used whereas in this study it was set to 2.0Å. That could be the reason why re-ranking did not have a significant effect.

In the experimental studies for solvent mapping, it was noted that favourable solvent positions on the protein surface appear at particular concentration of solvents in which crystal soaking experiments were performed[81, 82]. The concentration at which a solvent molecule is bound to a binding pocket in the protein, therefore, can be considered as a rough measure of the strength of binding. Similarly each MCSS pose has an associated score which also reflects binding affinity. We therefore compared the score of the experimentally nearest pose for each solvent position with its concentration. It should be noted that several solvent positions appear at the same concentration reflecting that those positions are identically favourable. The data for concentrations was obtained from the literature, and the resulting plot is shown in Figure 4.13.

A strong negative correlation should indicate good agreement as the MCSS score for a solvent pose observed at low concentration should be high. It was noted that the correlation coefficient was very weak (Figure 4.13). This is consistent with some of the limitations of MCSS scoring that were observed in the analysis of predicted solvent poses. It was however noted that the solvent molecules binding at high concentrations (≥ 80%) were clustered mostly below an MCSS score of 15.00 (Figure 4.13). At least three solvent probes binding at very low concentrations (DMF1004, 40%, IPA1, 10% and IPH1, 0.4%) were located outside this cluster with relatively high MCSS score. IPH1 position which was predicted with a very low $RMSD_{X\text{-}ray}$ and was rank 2nd but also obtained a high MCSS score. It was therefore concluded that rank ordering resulting from MCSS for solvent mapping datasets was incorrect in most cases (Table 4.8 and 5.4). Experimentally relevant positions were, however, generated for each solvent probe and scoring efficiency was relatively better for

buried sub-sites such as S3 for elastase and S1', where solvent screening of electrostatic interactions are less significant.

Protein	Solvent	MCSS Score	Conc. (%v/v)[81, 82]
ELS	ACN1001	10.87	95
	IPA1001	13.45	80
	IPA1002	14.62	40
	IPA1003	17.86	80
	DMF1004	19.96	55
	ETH1001	15.17	80
	ETH1002	13.18	80
	ETH1003	15.06	80
	ETH1004	9.99	80
	HEX1001	7.12	80
	HEX1004	8.25	80
	TFE1001	15.13	40
	TFE1002	9.74	40
	TFE1003	13.57	40
	TFE1008	9.69	40
TLN	ACN1	26.93	70
	CCN1	23.18	80
	IPH1	22.38	0.4
	IPA1	15.05	2
	IPA5	13.41	90
	IPA8	14.76	90
	IPA9	12.57	100

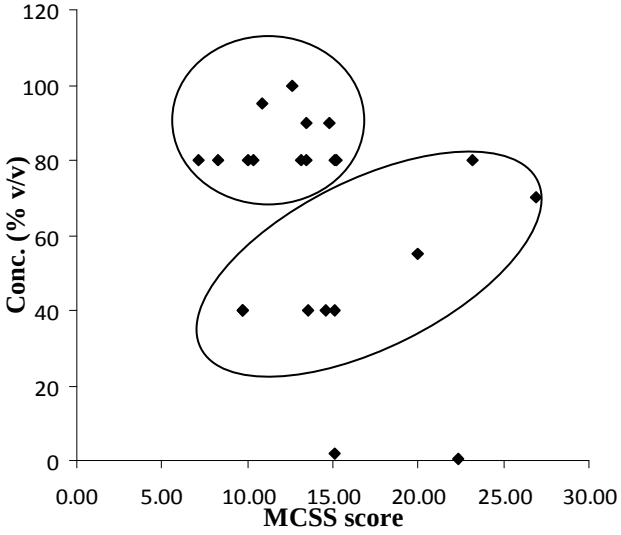


Figure 4.13. Correlation between MCSS score and concentrations of solvent probes from experimental mapping studies. The table shows solvent positions and the concentration at which those were observed in the binding site. The plot shows a weak negative correlation with R = -2.01.

4.11 Summary

The results of MCSS calculations on solvent mapping datasets indicate that MCSS can generate experimentally relevant poses for solvent probes. However, the ranking of poses is problematic and leads to false positive energy minima. A part of this problem, as highlighted before, is caused by the crude approximation of solvent

effects using a distance-dependent dielectric model. As a result, charge interactions between two polar groups that are considered favourable by the force-field could, in fact, be unfavourable[76]. The over-estimation of electrostatic interactions and appropriate treatment of solvent screening effects poses a significant challenge in protein-ligand docking and scoring[53]. In principle this should be overcome by an explicit treatment of solvent molecules during binding affinity prediction. Explicit solvent treatment through molecular dynamics simulation is computationally very expensive and might not be feasible at a large scale[35]. Over the past few years, however, advances have been made in the use of implicit solvent models based on continuum electrostatics[50]. These approaches treat solvent as a continuum of high dielectric medium whereas the protein is treated as a low dielectric medium with immersed partial charges.

Chapter 5

Fragment Docking and Scoring with MCSS and MM/GBSA Rescoring

5.1 Introduction

The results of MCSS calculations on solvent mapping datasets indicate that MCSS can generate experimentally relevant poses for solvent probes. However, the ranking of poses is problematic and leads to false positive energy minima. A part of this problem, as highlighted before, is caused by the crude approximation of solvent effects using a distance-dependent dielectric model. As a result, charge interactions between two polar groups that are considered favourable by the force-field could, in fact, be unfavourable[76]. The over-estimation of electrostatic interactions and appropriate treatment of solvent screening effects poses a significant challenge in protein-ligand docking and scoring[53]. In principle this should be overcome by an explicit treatment of solvent molecules during binding affinity prediction. Explicit solvent treatment through molecular dynamics simulation is computationally very expensive and might not be feasible at a large scale[35]. Over the past few years, however, advances have been made in the use of implicit solvent models based on continuum electrostatics[50]. These approaches treat solvent as a continuum of high dielectric medium whereas the protein is treated as a low dielectric medium with immersed partial charges.

Fragment docking and scoring programs that make use of implicit solvent models have been discussed before. One of the examples reported earlier was based on

combining MCSS with an MM-PB/SA re-scoring scheme[74]. A significant re-ordering of MCSS poses was reported with this approach. As PBSA is considered the most rigorous and time-consuming implicit solvent treatment, faster approximations to PBSA have been developed, such as the generalized Born (GB) model, which in some cases, reproduces results from PBSA at error margin of ≤ 1%[101].

Therefore, we attempted to use a GB implicit model in the GBSA re-scoring scheme along with MCSS. This particular GB model uses molecular volume integration for efficient approximation of molecular surface area calculation, hence it is called GBMV/SA[186]. The details of this scoring strategy are explained in Chapter 4 (4.4.6). The results of the application of these calculations on fragment docking and HSP90 datasets are described below.

5.2 MCSS calculations on fragment docking dataset

The result of the MCSS calculations is a set of poses for each fragment in a binding site, each with an associated score calculated within the MCSS protocol (a CHARMm interaction energy) and with a subsequent MM-GBMV/SA rescoring. The best scoring post for each fragment was compared with the original X-ray pose and the *in situ* minimized X-ray pose and the success rate assessed at 1 Å and 2 Å RMSD cut-offs. As *in situ* minimization of the X-ray pose sometimes leads to significant deviations therefore it is important to consider both poses as references. Table 5.1 summarizes the results for the fragment docking dataset. The success rate is defined as the percentage of fragments for which the top-scoring pose was predicted within the RMSD cut-off. It was noticed that the success rate for MCSS was the same at 1Å and 2Å RMSD regardless of the reference pose. The same trend was observed after

rescoring with MM-GBMV/SA. As expected, the rescoring improved the success rate up to 67% at 1Å RMSD and 75% at 2Å RMSD.

Table 5.1. Success rate of MCSS and MCSS with MM-GBMV/SA scoring on fragment docking dataset, at different RMSD cut-offs and considering X-ray and *in stiu* minimized X-ray poses as reference.

RMSD	**1Å**		**2Å**	
Reference	**X-ray**	**X-ray Min**	**X-ray**	**X-ray Min**
MCSS	50%	50%	67%	67%
MM-GBMV/SA	67%	67%	75%	75%

Further details of results from MCSS are summarised in Table 5.2 where in each case, the top-scoring pose from MCSS is shown, along with its RMSD from the X-ray pose ($RMSD_{X\text{-}ray)}$ and from the *in situ* minimized X-ray pose ($RMSD_{X\text{-}rayMin}$). The RMSD between the X-ray pose and the *in situ* minimized X-ray pose is also shown ($RMSD_{Xray|XrayMin}$). These values lie in the range 0.2Å to 1.0Å. High $RMSD_{Xray|XrayMin}$ values (> 0.50 Å) are mostly associated with resolution lower than 2.0 Å, as expected. Further inspection of the case, for which an MCSS top-scoring pose was not within 1.0Å or 2.0Å of any of the reference poses, revealed that the MCSS protocol had found a correct pose in all cases, but the MCSS score was not able to identify it reliably. The results after rescoring with MM-GBMV/SA are summarised in Table 5. For each case, the most favourable pose obtained after rescoring is shown along with RMSDs against references. The success rate increases with improvement in scoring. This was particularly noticeable for 1WCC, 2C90 and 2JJC where poses within 1.0Å of the references pose were picked up by MM-GBMV/SA scoring as the most favourable ones.

Table 5.2. Results of MCSS docking for fragments in fragment docking dataset (top-scoring poses for each test case). ($RMSD_{Xray|Xray\ Min}$: RMSD of *in situ* minimized X-ray pose from X-ray pose, $RMSD_{X\text{-}ray}$: RMSD of the best pose from the X-ray pose, $RMSD_{X\text{-}rayMin}$: RMSD of the best pose from the *in situ* minimized X-ray pose).

PDB	$RMSD_{Xray\|XrayMin}$	$RMSD_{X\text{-}ray}$	$RMSD_{X\text{-}ray\ Min}$
1EQG	0.3	0.3	0.1
1FV9	1.0	2.0	1.9
1GWQ	0.5	1.5	1.5
1N1M	0.5	0.8	0.4
1QWC	0.8	1.0	1.2
1S39	0.3	0.3	0.0
1WWC	0.2	4.8	4.8
1YZ3	0.4	0.4	0.0
2ADU	0.5	0.5	0.1
2C90	0.6	5.2	5.0
2JJC	0.4	6.3	6.4
2OHK	0.5	2.7	2.6

In total, there were three failures: 1GWQ, 2ADU and 2OHK. Further analysis revealed that 1GWQ was scored correctly by MCSS alone with 1.5 Å $RMSD_{X\text{-}ray}$ of the top-scoring pose but after re-scoring, this pose obtained a ΔG of binding that was only 0.7 kcal/mol higher than that of the most favourable pose with 7.1 Å $RMSD_{X\text{-}ray}$. The ΔG values for X-ray, *in situ* minimized X-ray and top-scoring poses is given in Appendix. These two poses differ in the orientation of two distal phenolic groups in the fragment. The opposite orientation of these groups still satisfies their hydrogen bonding potential but the position of sulphur atom in the thiophene ring is different which could explain the minor difference in the binding energy values of the two poses (Figure 5.1).

In the case of 2OHK the top ranking pose from MCSS remained the most favourable binding pose with RMSDX-ray and $RMSD_{X\text{-}rayMin}$ values 2.6Å and 2.7Å, respectively (Table 5.3). On the other hand, another MCSS pose at $RMSD_{X\text{-}ray}$ 1.26Å and $RMSD_{X\text{-}}$

$_{rayMin}$ 1.04Å improved from rank 4 to rank 3 but did not achieve the most favourable binding energy.

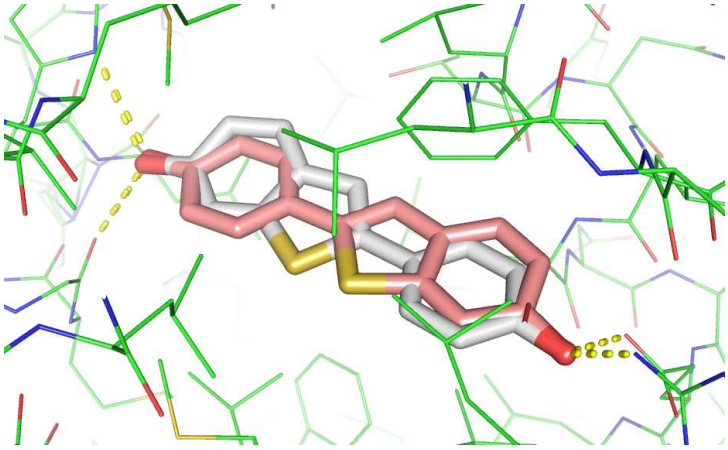


Figure 5.1. MCSS generated poses for 1GWQ ranked 1st (magenta C-atoms) and 2nd (grey C-atoms) by MM/GBMV-SA. The 2nd pose which corresponds to the X-ray pose (RMSD 1.5Å) is 0.6 kcal/mol less favourable than the 1st pose which is in completely opposite orientation to the X-ray pose with high RMSD (7.1Å).

Table 5.3. The most favourable MCSS poses after MM-GBMV/SA scoring for fragments in fragment docking dataset. ($RMSD_{X\text{-}ray}$: RMSD of the nearest pose from the X-ray pose, $RMSD_{X\text{-}ray\ Min}$: RMSD of the nearest pose from the *in situ* minimized X-ray pose).

PDB	$RMSD_{X\text{-}ray}$	$RMSD_{X\text{-}ray\ Min}$
1EQG	0.3	0.1
1FV9	2.0	1.9
1GWQ	7.1	7.1
1N1M	0.8	0.4
1QWC	1.0	1.2
1S39	0.3	0.0
1WWC	0.2	0.0
1YZ3	0.4	0.0
2ADU	9.1	9.1
2C90	0.6	0.0
2JJC	0.3	0.0
2OHK	2.7	2.6

2ADU is a unique entry in the dataset as the fragment interacts with two Co^{2+} ions in the binding site. The parameters for Co^{2+} ions were missing in the MMFF94 force-field so these were changed to Zn^{2+} ions for calculations. The resulting ΔG values obtained for the X-ray and *in situ* minimized X-ray poses were all positive. Similarly, most of the MCSS poses obtained a positive binding energy value. The most favourable pose was located very far from the binding site. Interestingly, MCSS produced a correct solution but the re-scoring reduced the quality of prediction, giving the most favourable energy to a different pose that was located very far from the X-ray pose (Table 5.3). It was noticed that all MCSS poses which were interacting with the metal ions received unfavourable binding energy values like the X-ray and X-ray minimized pose.

This can probably be explained by the lack of adequate parameterization of interactions involving metal ions. This was also noted for 1QWC where the X-ray pose of the fragment makes key interactions with the heme group in the binding site. The amidine group stacks on top of heme pyrrole B with hydrogen bonding to Glu592 and 3-aminomethyl group also makes salt bridges with heme propionic acid groups[187]. The candidate poses generated from both MCSS protocols were dominated by distorted fragment conformations where either amidine nitrogen atom or amino group nitrogen atom interacted with iron atom at the heme centre, resulting in unrealistic binding energy values after rescoring. Only after discarding these poses was a correct ranking was obtained (Table 5.3).

As MCSS is an energy minimization routine, the efficiency of sampling low energy conformations of fragments in the binding sites is an important consideration. One way of assessing this is by plotting the difference in ΔG values obtained for the top-

scoring (lowest energy pose) and the *in situ* minimized X-ray pose. This could indicate if MCSS minimization is finding energy minima in the binding site that would otherwise be found by minimizing the X-ray pose in the binding pose.

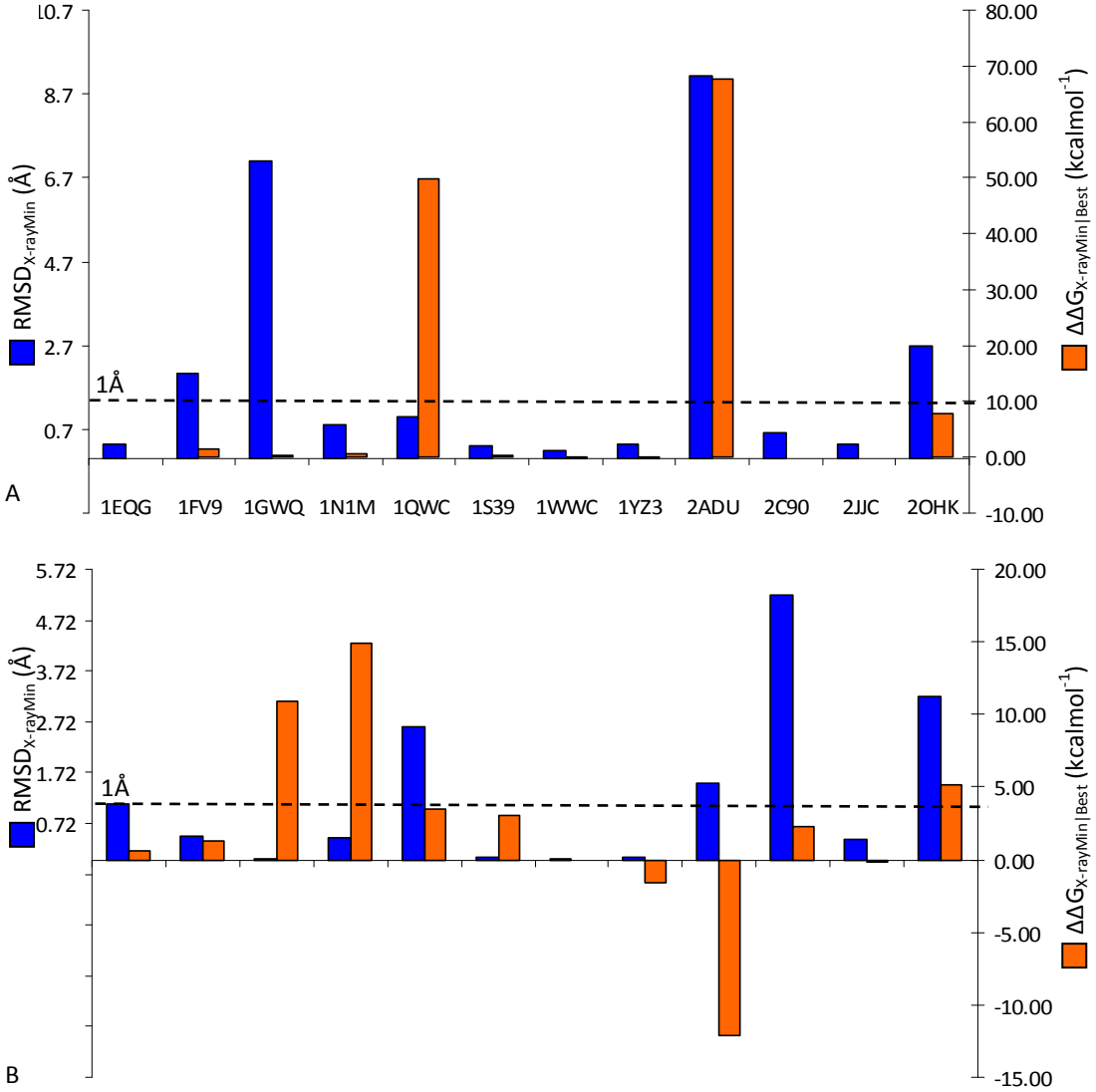


Figure 5.2. Most favourable MCSS and GOLD poses after MM-GBMV/SA scoring in fragment docking dataset. The difference in binding affinity with the in X-ray situ minimized pose ($\Delta\Delta G_{X\text{-}rayMin|Best}$) is plotted alongside RMSD with respect to *in situ* minimized pose ($RMSD_{X\text{-}rayMin}$) for each case.

The values for ΔG calculated for the X-ray pose, *in situ* minimized X-ray pose and top-scoring poses are shown in the Appendix (Table 7.3 and 7.4). Here only differences in

ΔG values are described as they are more meaningful. Figure 5.2 shows plots for the dataset where the difference in ΔG values ($\Delta\Delta G_{X\text{-}rayMin|Best}$) is plotted alongside the $RMSD_{X\text{-}rayMin}$ of the top-scoring pose. A positive value of $\Delta\Delta G_{X\text{-}rayMin|Best}$ reflects that docking produced a low energy pose. The RMSD bar next to $\Delta\Delta G_{X\text{-}rayMin|Best}$ indicates if this low energy pose corresponds to the *in situ* minimized X-ray pose. In this way, it is possible to see if incorrect ranking is related to poor sampling efficiency. The plot for fragment docking dataset shows that in almost all cases docking sufficiently sampled the low energy conformations (Figure 5.2A). This indicates that MCSS consistently finds low energy poses in the binding site therefore it is mostly the imperfections in scoring that lead to false positives.

There are two cases that lie between the 1.0Å and 2.0Å RMSD cut-off (Table 5.3). 1FV9 represents a case where the top-scoring pose is about 2.0Å away from the reference poses. Although for the most part the two poses superpose nicely, the top-scoring pose is inverted with reference to the position of hydroxyl group attached to the benzene ring (Figure 5.3). This highlights the importance of using rather strict criterion when assessing the performance of scoring functions on protein-fragment complexes.

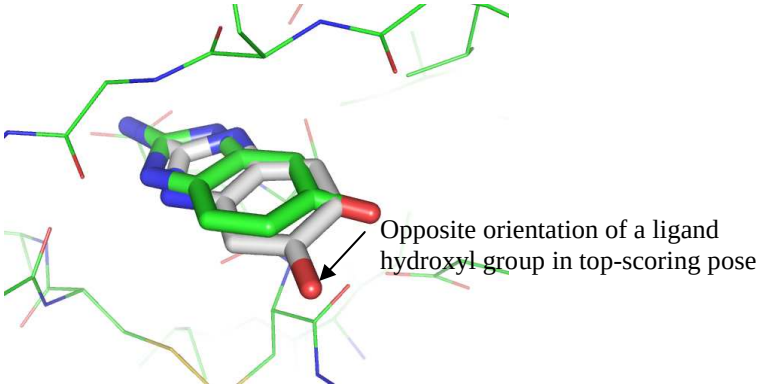


Figure 5.3. Top-scoring MCSS pose for 1FV9 at 2.02 Å RMSD. The top-scoring pose (grey C-atoms) overlaps with the X-ray pose (green C-atoms) for the most part, except a hydroxyl group.

5.3 Comparison with GOLD

In order to compare the performance of MCSS and MM-GBMV/SA scoring, the same procedure was repeated with a standard docking program, GOLD, in the place of MCSS. Table 5.4 shows the success rate obtained from GOLD at 1.0 Å and 2.0 Å RMSD cut-offs with respect to X-ray and *in situ* minimized X-ray poses. The performance of GOLD docking and scoring at 1.0 Å with either of the two references was slightly better than that of MCSS. At 2.0 Å the success rate was similar to MCSS (Table 5.1 and 5.4). The rescoring of GOLD poses with MM-GBMV/SA resulted in a success rate at 1.0 Å that was lower than equivalent success rate for MCSS. At 2.0 Å, however, the success rate after rescoring after either of the two methods is the same (Table 5.1 and 5.4).

Table 5.4. Success rate of GOLD and GOLD with MM-GBMV/SA scoring on fragment docking dataset, at different RMSD cut-offs and considering X-ray and *in situ* minimized X-ray poses as references.

RMSD	**1Å**		**2Å**	
Reference	**X-ray**	**X-ray Min.**	**X-ray**	**X-ray Min.**
GOLD	58%	58%	67%	67%
GOLD-GBMV/SA	50%	58%	75%	75%

Further analysis of the results from GOLD docking is summarized in Table 5.5. Although the top-scoring pose in 5 cases did not correspond to the X-ray or *in situ* minimized X-ray pose, the docking search yielded a pose within 2.0 Å of either of the references for all cases. The effect of rescoring was therefore further investigated to understand the improvement in ranking of these cases.

Further analysis of the results from MM-GBMV/SA rescoring of GOLD poses is summarized in Table 5.5 and 5.6. The increase in the success rate after rescoring was

due to the ranking of correct poses for 1FV9 and 1WWC. In three other cases, 1QWC, 2C90, 2OHK, rescoring did not improve results.

The plot of $\Delta\Delta G_{X\text{-}rayMin|Best}$ and $RMSD_{X\text{-}rayMin}$ for the top-scoring poses after rescoring shows that failures in scoring were not related to an inadequate search for low energy conformers (Figure 5.2B) and GOLD found correct poses for almost all cases.

Although the performance of MCSS and GOLD after rescoring is comparable, it can be seen from Figure 5.2A and B that most of the failures are different in each method. The combined result gives success rate of 75% at 1Å RMSD considering *in situ* minimized X-ray poses as the reference.

Table 5.5. Results of GOLD docking for fragments in fragment docking dataset (top-scoring poses for each test case). ($RMSD_{Xray|XrayMin}$: RMSD of *in situ* minimized X-ray pose from X-ray pose, $RMSD_{X\text{-}ray}$: RMSD of the best pose from the X-ray pose, $RMSD_{X\text{-}rayMin}$: RMSD of the best pose from the *in situ* minimized X-ray pose).

PDB	$RMSD_{X\text{-}ray\|XrayMin}$	$RMSD_{X\text{-}ray}$	$RMSD_{X\text{-}rayMin}$
1EQG	0.3	0.3	0.3
1FV9	1.0	2.0	1.9
1GWQ	0.5	0.7	0.1
1N1M	0.5	0.5	0.8
1QWC	0.8	5.2	5.2
1S39	0.3	0.2	0.4
1WWC	0.2	6.2	6.2
1YZ3	0.4	0.7	0.9
2ADU	0.5	0.6	0.5
2C90	0.6	5.2	5.0
2JJC	0.4	0.3	0.4
2OHK	0.5	3.2	3.3

Table 5.6. The most favourable GOLD poses after GBMV scoring for protein-fragment complexes from fragment docking dataset.

PDB	$RMSD_{X\text{-}ray}$	$RMSD_{X\text{-}rayMin}$
1EQG	1.2	1.1
1FV9	1.2	0.5
1GWQ	0.6	0.0
1N1M	0.8	0.4
1QWC	2.5	2.6
1S39	0.4	0.1
1WWC	0.2	0.0
1YZ3	0.4	0.1
2ADU	1.3	1.5
2C90	5.1	5.2
2JJC	0.3	0.4
2OHK	3.3	3.2

5.4 MCSS-GBMV calculations on HSP90 dataset

The docking and scoring protocol with MCSS and MM-GBMV/SA was applied to the HSP90 dataset which contains 11 HSP90 N-terminal domain structures bound mostly to fragment sized molecules. All the fragments were docked into their native and the non-native receptor structures in the dataset. Table 5.7 and 5.8 summarize the results obtained from docking fragments into their native structures with MCSS and scoring based MM-GBMV/SA scoring for each case.

Table 5.7. Success rate of GOLD and GOLD with MM-GBMV/SA scoring on HSP90 dataset, at different RMSD cut-offs and considering X-ray and *in situ* minimized X-ray poses as references.

RMSD	1Å		2Å	
Reference	**X-ray**	**X-ray Min.**	**X-ray**	**X-ray Min.**
GOLD	17%	25%	25%	25%
GOLD-GBMV/SA	50%	58%	58%	67%

The success rate for docking into the native receptor structures was 58% and 67% at 1Å and 2Å RMSD, respectively, using the *in situ* minimized X-ray pose as the reference (Table 5.7). This was significantly better than original success rate from MCSS before re-scoring.

It can be seen from Table 5.8 that for 7 cases, the predicted pose with the lowest $RMSD_{X\text{-}ray}$ was assigned the most favourable binding energy. There were five cases where the most favourable binding pose was far from the X-ray pose. Further analysis indicated that for two of such cases, 1QYE and 2QFOa, the lowest $RMSD_{X\text{-}ray}$ pose was ranked 2nd whereas in other three cases, 2CCS, 2QFOb and 3EKO, the lowest $RMSD_{X\text{-}ray}$ pose was ranked 6th, 7th and 4th respectively. These observations are reflected in the plot of $\Delta\Delta G_{X\text{-}rayMin|Best}$ and $RMSD_{X\text{-}rayMin}$ which shows the MCSS efficiently samples low energy conformations in most cases and mostly it is the scoring function that is unable to assign the most favourable binding energy to the energy minima that is closest to the X-ray or *in situ* minimized pose (Figure 5.4).

Table 5.8. Most favourable MCSS poses after docking and MM-GBMV/SA scoring of fragments in native receptor structures from HSP90 dataset.

PDB	$RMSD_{X\text{-}ray}$	$RMSD_{X\text{-}rayMin}$
1QYE	4.10	4.13
1ZWH	2.07	1.08
2CCS	5.91	5.89
2JJC	0.34	0.03
2QF6	0.36	0.13
2QFOa	3.72	3.61
2QFOb	2.29	2.08
2WI1	0.92	0.11
2WI2	1.00	0.90
3BM9	0.32	0.62
3EKO	5.06	4.95
3FT5	0.78	0.71

The values for ΔG calculated for the X-ray pose, *in situ* minimized X-ray pose and top-scoring poses are shown in the Appendix (Table 7.5)

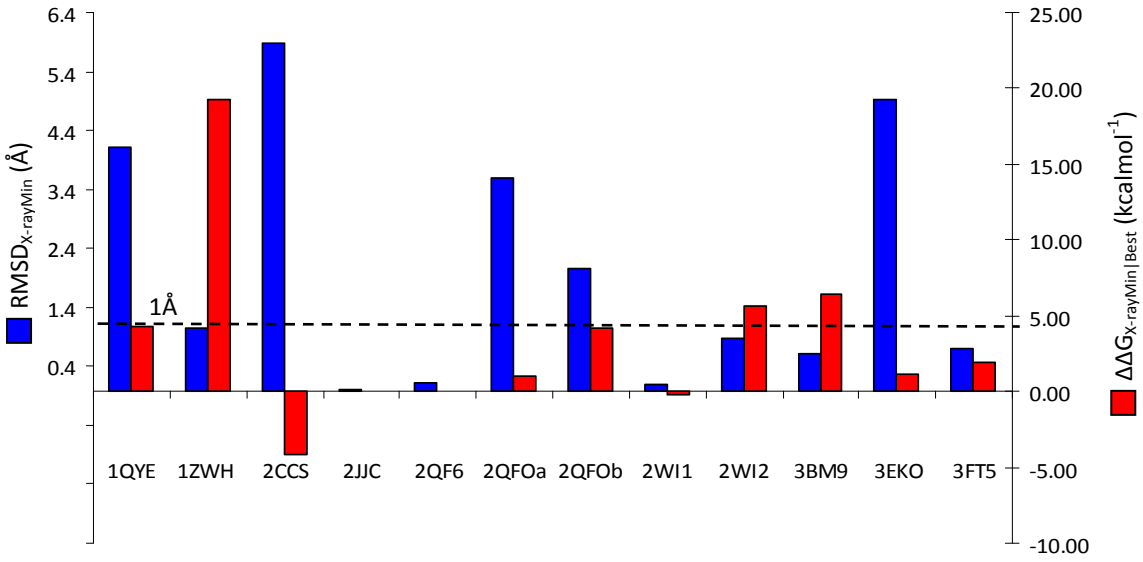


Figure 5.4. Most favourable MCSS and GOLD poses after MM-GBMV/SA scoring in HSP90 dataset. The difference in binding affinity with the *in situ* minimized pose ($\Delta\Delta G_{X\text{-}rayMin|Best}$) is plotted alongside RMSD with respect to *in situ* minimized pose ($RMSD_{X\text{-}rayMin}$) for each case.

A distinct case in HSP90 dataset is that of 1ZWH where the most favourable pose is on the margin of the 2.0Å threshold (Table 5.8). This is also the case with the largest deviation between the X-ray pose and *in situ* minimized pose. In the X-ray binding mode of the fragment, the carboxylic group attached to the ring is partially exposed to the solvent. The *in situ* minimization which was done with distance-dependent dielectric model favoured a binding mode where this group formed a hydrogen bond with ASN92 in the binding site. It should be noted that the MCSS minimization is also carried out with distance-dependent dielectric. The initial placement of fragments

can therefore result in an experimentally relevant pose getting a very low MCSS rank. In the case of 1ZWH, however, an MCSS pose which reproduced most of the key interactions in the native binding mode obtained the most favourable binding energy after re-scoring.

2QFOa and 2QFOb represent a case of co-operative binding. It was shown that 2QFOb binds only in the presence of 2QFOa[43]. The native binding mode of these fragments also indicates π-π stacking interactions between the pyrimidine ring of 2QFOa and the phenyl ring of 2QFOb. In this protocol, these fragments were docked and scored independent of each other therefore energetic contributions resulting from the direct interactions between these two fragments could not be accounted for. This probably contributed to the lowest $RMSD_{X\text{-}ray}$ poses for 2QFOa and 2QFOb getting rank 2 and 7, respectively. The difference in the top scoring and lowest $RMSD_{X\text{-}ray}$ poses for 2QFOa was only about 0.5 kcal/mol. In the top scoring pose of 2QFOb, the furanone moiety was accurately superimposed on the X-ray pose and the major deviation came from the benzene ring which is stabilized by staking interactions with pyrimidine ring of 2QFOa (not present in 2QFOb docking and scoring).

The most compelling reason for using MM-GBMV/SA re-scoring of MCSS poses is the treatment of solvent effects that take into account screening of electrostatic interactions between polar groups. Therefore, the initial MCSS ranking of lowest $RMSD_{X\text{-}ray}$ poses is expected to be improved in GBMV scoring, in at least those cases where electrostatic interactions play major roles in binding affinity, as it was noticed previously that for 7 correct predictions (Table 5.8), the ranks of the lowest $RMSD_{X\text{-}ray}$ poses during MCSS improved or remained unchanged (3 cases). Similarly, in those

cases where the lowest $RMSD_{X\text{-}ray}$ pose did not obtain the most favourable binding energy, there was still considerable improvement in the ranking.

5.5 Effect of Multiple Receptor Structures

The HSP90 binding site is known for ligand induced conformational changes[175]. The base of the pocket (containing the purine binding site with conserved solvent structure binding to Asp93) remains relatively static across all known structures. The main change is the rearrangement of a helix at the lip of the binding site which can sometimes open up an additional hydrophobic pocket[42]. It is therefore important in docking and scoring of fragments to use the most appropriate target structure. One approach is to use multiple receptor structures obtained either from molecular dynamics simulation or from X-ray crystallography[36]. The fragments in the HSP90 dataset were therefore docked into the set of non-native structures to study how the consideration of multiple structures could possibly affect the outcome of docking.

Cross-docking indicated that the performance of individual receptor structures, in terms of percentage of fragments in the dataset correctly docked and scored varied from 0 to 45%. It was noticed that the correct binding mode for each fragment was reproduced and scored at the top in at least one or more receptor structures, except 2QFOb. The successive addition of structures can therefore increase the total performance up to 91%. Figure 5.5 shows the increase in the performance as structures are added randomly. The average performance of 100 random selections is plotted against the number of receptor structures and the error bars represent standard deviation in the average performance. As expected, the average

performance increased almost linearly with successive addition of structures at random, leading up to the performance of 91% when using all structures in the dataset. The increases in performance is observed at 1Å and 2Å RMSD criteria for success rate (Figure 5.5)

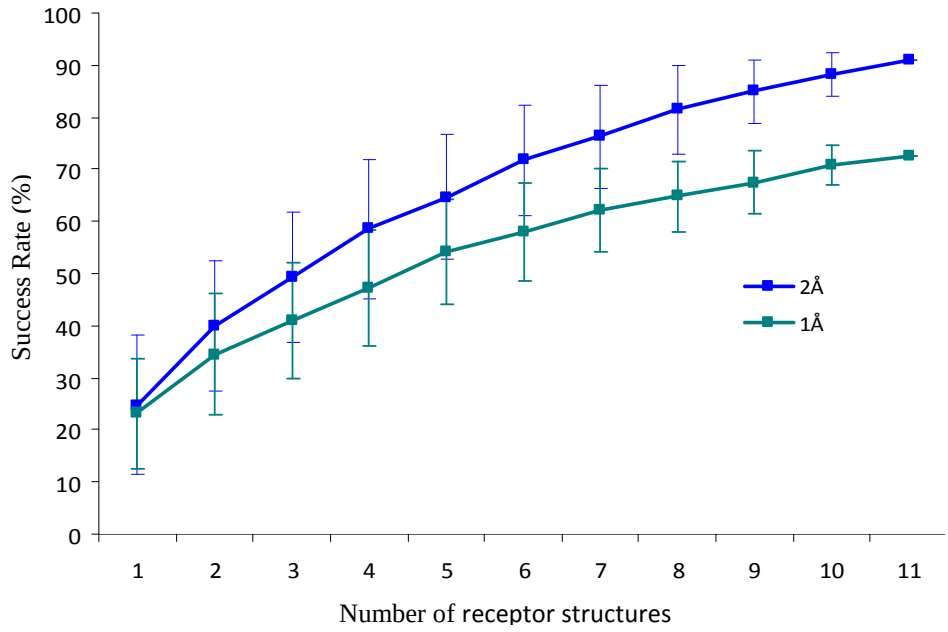


Figure 5.5. The effect of multiple structures on scoring performance. The scoring is defined as the percentage of correctly docked and scored ligands in at least one receptor structure. Successive addition of structures is based on 100 random selections at each point and the average performance is plotted against the number of structures with error bars representing standard deviation of success rate in 100 random selections of additional structures.

This performance, is however, based on the criterion that the X-ray crystallographic binding mode of the fragment is reproduced and scored correctly in at least one receptor structure. For brevity, the terms native and non-native receptor would imply the receptor structure in which fragments were docked. Different receptor structures in the dataset can either improve or deteriorate the quality of predictions.

If a native receptor predicts the X-ray binding mode with the most favourable interaction energy (the highest rank), it is possible that a non-native receptor gives poor rank to the lowest RMSD pose, thereby deteriorating the quality. Similarly, if a native receptor fails to reproduce and correctly rank the X-ray binding mode, a non-native receptor can improve the performance by reproducing the correct binding mode with the top rank.

In order to compare how non-native receptors perform with respect to the native receptor, the criteria were set for the improvement or deterioration of the quality of predictions as follows. The receptor structure is considered deteriorating when a native receptor reproduces a pose with ≤ 2.0Å $RMSD_{X\text{-}ray}$ with the highest score and the non-native receptor fails to score a pose with ≤ 2.0Å $RMSD_{X\text{-}ray}$ within top 3. The receptor structure is considered improving when a native receptor fails to reproduce the native binding mode and score it correctly and a non-native receptor reproduces a pose with ≤ 2.0Å $RMSD_{X\text{-}ray}$ within the top 3 poses. In this way for each fragment, the number of non-native receptors that have an improving or deteriorating effect on the quality of predictions can be calculated. This is shown in Figure 5.6 where for each fragment the number of improving versus deteriorating receptor structures is shown.

It was previously observed for the HSP90 dataset that in 5 cases prediction in the native structures was not successful (Table 5.8). It can be seen from Figure 5.6 that for such cases (1QYE, 2CCS, 2QFOa, 2QFOb and 3EKO) at least one non-native receptor was able to reproduce the native binding mode with correct scoring, thereby improving the quality of predictions. This was particularly notable for 2QFOa for which all of the non-native receptors reproduced the native binding mode.

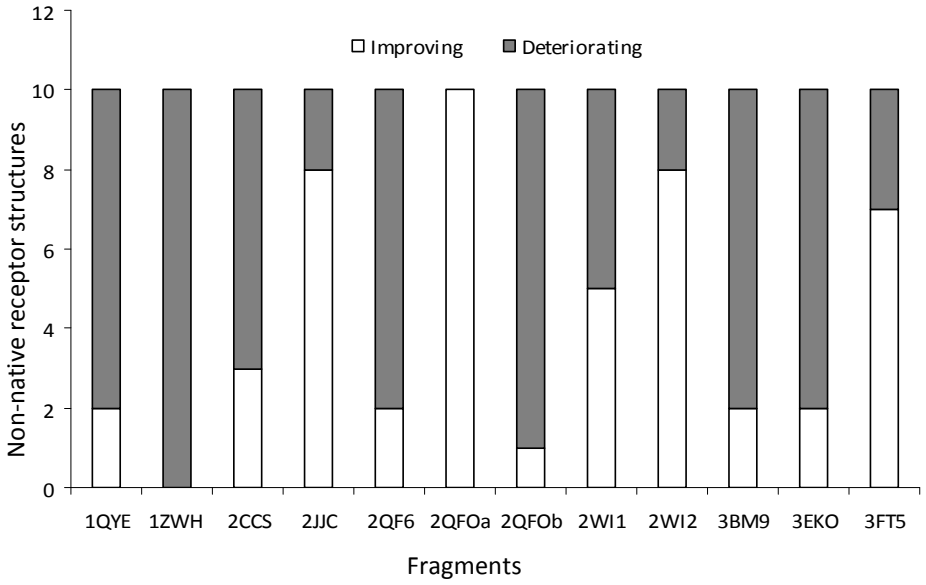


Figure 5.6. Relative contribution of non-native receptor structures towards scoring of each docked fragment in HSP90 dataset.

For the 7 cases where docking and scoring in native receptor was successful (1ZWH, 2JJC, 2QF6, 2WI1, 2WI2, 3BM9 and 3FT5), a fraction of non-native receptors deteriorating the quality of scoring and ranking were also present (Figure 5.6). The most significant of these was 1ZWH where none of the non-native receptors could reproduce the native binding mode with correct ranking.

This suggests that it can be useful to assess relative contribution of multiple structures in docking and scoring known ligands before using them in a library screen. The balance between the improving and deteriorating effect of multiple structures on the quality of predictions should therefore be taken into account.

5.6 Prediction of conserved water molecules in the binding site

The role of conserved water molecules in modulating interactions in the HSP90 binding site is well established[44]. The treatment of these water molecules as part of the binding site contributed to reasonable predictions in docking and scoring. An

important question therefore is how well MCSS can predict the positions of ordered water molecules in the binding site. The HSP90 binding site can be used as a test case.

In this dataset 3 to 4 interstitial water molecules were observed in the binding site, mostly bridging interactions between protein and the bound ligand. These are referred to as HOH1, HOH2, HOH3 and HOH4 (Table 5.9) (Figure 5.7). Apart from HOH2, all of these water molecules were present in all 11 structures whereas HOH2 was present only in 7 structures.

In order to predict the position of water molecules, the MCSS protocol mentioned above was applied to all receptors in HSP90 dataset with water as the probe. It was noticed that the top ranking predicted water position in all receptor structure always corresponded to the crystallographic water position HOH1, with an average RMSD of 0.27Å. The water position HOH4 was predicted with an average RMSD of 0.56Å and was ranked variably from 2 to 4 in 9 structures. Similarly, HOH3 was predicted with an average RMSD of 0.49Å and was ranked variably from 3 to 5 in 8 structures. Finally, HOH2 was predicted only in one of 7 structures where it is present. It should be noted that the position of HOH2 in HSP90-ligand complexes is replaced by an oxygen atom from the ligand in some cases (Figure 5.7). The displacement of HOH2 by ligand indicates that it is relatively less stabilized as compared to other water positions. The prediction of conserved water molecules in HSP90 binding site was quite reasonable. A detailed study on the prediction of ordered water molecules in different binding site using MCSS would be required to further validate these results.

Table 5.9. Conservation of four key water positions across multiple HSP90 structures. The residue names of water molecules in PDB files corresponding to each of the positions are shown.

PDB	HOH1	HOH2	HOH3	HOH4
1QYE	W4	W19	W43	W68
1ZWH	W3	--	W5	W4
2CCS	W174	--	W303	W173
2JJC	W166	W292	W78	W164
2QFO	W12	W3	W11	W6
2WI1	W94	W123	W33	W59
2WI2	W86	W119	W23	W48
3BM9	W8	--	W2	W4
3EKO	W1	--	W3	W2
3FT5	W11	W5	W14	W15

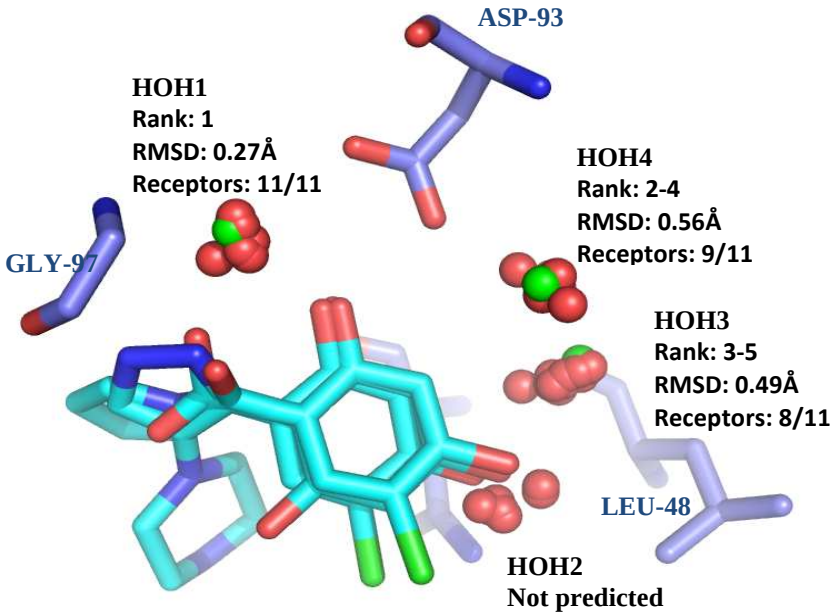


Figure 5.7. Prediction of conserved water molecules in HSP90 binding site using MCSS. Crystallographic position of water molecules in the 11 structures in HSP90 dataset (red) are superposed along with the bound ligand. The averaged predicted position from all receptor structures is shown in green. For each water position, the following details are shown: the rank of nearest predicted pose, its average RMSD based on prediction in all structures, the number of receptor structures where it was predicted versus total number of receptor structures where this water position was observed.

5.7 Discussion

MCSS is one of the longest established methods designed with the aim of predicting energetically favourable positions of functional groups in protein binding sites[69]. As the fragment-based lead discovery methods come of age[60], the significance of computational techniques to aid the discovery of novel fragments, that could be used as starting points for designing potent lead compounds, has increased tremendously[67]. The challenges posed by docking and scoring of fragments persist because of the promiscuity of fragment binding and the limitations of scoring functions. In the context of MCSS, a recent review[76] highlighted the need for 1) evaluation of MCSS minima and their ranking for high quality predictions and 2) sampling multiple receptor conformations as receptor is held rigid in MCSS minimization. In this study we investigated both of these issues by applying MCSS calculations followed by binding energy estimation on a set of 12 different protein-fragment complexes and on a set of 11 flexible HSP90 protein- fragment complexes.

In the MCSS scoring function the solvent effects are taken into account by a very crude approximation of dielectric screening from a distance-dependent dielectric model. This has led to inaccuracies in scoring as the desolvation penalty accompanying some of the polar interactions between the protein and the functional group is not appropriately taken into account. An alternative approach is to couple the molecular mechanical energy component with polar and non-polar solvation free energies derived from implicit solvent formalisms in MM-GB/SA or PB/SA method. The sorting of MCSS minima based on MM-PB/SA derived free energy estimates was previously described and showed encouraging results. The MM-PB/SA method, considered as benchmark in implicit solvent methods, is time-

consuming and less feasible on a large-scale. Continuing developments in implicit solvent methods have led to semi-analytical solutions to PB equation with high accuracy, such as the GB equation[107]. It is therefore timely to investigate and evaluate the performance of MCSS combined with a MM-GB/SA as a fragment docking and scoring scheme. The calculations based on a set of protein-fragment complexes suggest encouraging results with some limitations.

The assessment of the success rate was complicated by factors such as the choice of reference pose and the RMSD cut-off for comparison with the predicted poses. As MCSS is an energy minimization routine so the X-ray position for each fragment was *in situ* minimized to provide an additional reference for the comparison. Similarly, due to the small size of fragment molecules the RMSD cut-off for determining the success rate was set at 1Å and 2Å to highlight borderline cases. It was also noticed that the choice of reference position was more relevant at lower RMSD cut-off as the success rate converged to similar values at 2Å regardless of the reference chosen.

Using MCSS primarily as a method to generate poses for subsequent evaluation with a more rigorous scoring function shows significant increase in scoring accuracy as compared to docking and scoring with MCSS alone. The observed increase in the percentage of correctly docked and scored ligands was from 50% to 75%. This was the same amount of improvement that was observed for a standard docking and scoring method, GOLD, combined with MM-GB/SA scoring. Interestingly, despite the similar success rate both methods show success and failure for different cases. This enables a combined success rate of 83% and suggests that probably docking and scoring by combining the output from different methods is a suitable approach.

Occasionally, MCSS was unable to generate a pose close to any of the reference poses which probably indicates incomplete sampling of low energy conformations in the binding site. In most failures however, non-experimentally relevant energy minima were given favourable binding energy values.

The main source of poor ranking is the incomprehensive treatment of solvent effects which is largely rectified in the GB/SA scoring step. This is reflected in the observed improvement of ranking. As with any other docking program sampling receptor conformations is a challenging aspect in MCSS. Multiple receptor structures from X-ray crystallography or molecular dynamics simulations provide a way of sampling different conformations[36]. Therefore, we studied the feasibility of using this docking and scoring scheme for a flexible target by including multiple ligand-bound conformations. It was noted that random addition of structures to see if a particular fragment can be docked and scored correctly in at least one receptor structures in the dataset increases the overall percentage performance. However, some receptor structures would improve the quality of predictions by reproducing correct binding mode and interaction energy where native receptor structure has failed to do so. Similarly, some receptor structures produce incorrect scoring following correct scoring in the native structure, thereby reducing the quality of prediction. It could probably be useful to benchmark a set of receptor structures based on how efficiently they reproduce and score binding modes of non-native ligand/fragments. This can then be used as a pre-selection criterion for how many and which receptor structures could be used for optimal results. It should be noted that the plastic nature of binding in flexible active sites cannot be fully accounted for in this

approach as the ligand could induce novel conformational variations in the binding site.

We also revisited the capability of MCSS in to predict energetically favourable binding positions of functional groups in the binding site. As the treatment of conserved water molecules as part of the binding site in HSP90 docking and scoring yielded very good quality predictions, we asked the question if the positions of these ordered water molecules could be predicted at the first place. For HSP90 binding site, the results show very good correlation with experimentally observed positions. The ranking of these positions, without solvent correction, also corresponded reasonably with some experimental observations such as the top ranking cluster of waters in all receptor structures always corresponded to an experimentally observed water position. Similarly, often a low ranking MCSS predicted cluster corresponded to a much less stabilized water molecule in the HSP90 binding site which is often displaced by the ligand. A detailed study on prediction of conserved water molecules in a variety of binding sites would be required to further investigate the consistency in these predictions.

The work described in this chapter has shown that MCSS followed by re-scoring with MM/GBSA is a more successful method than GOLD for predicting fragment position in a target protein site. However, without re-scoring, the results from MCSS were poorer than those obtained with GOLD. The success rate of MCSS plus MM-GB/SA is high enough that the technique could be a useful method to support early ligand discovery efforts when it is not possible to determine crystal structures of fragments binding to proteins.

The conformational variations in the receptor structure upon binding and water molecules at the binding interface pose significant challenges. We show that the use of multiple receptor structures can possibly increase the success rate but at the expense of including receptor structures that also deteriorate quality of predictions. For HSP90, reasonable prediction of highly conserved water molecules in the binding site were obtained with MCSS but its applicability a large and more challenging scenario is yet to be evaluated.

6. Concluding Remarks

This thesis describes the characterisation of protein-ligand interactions using cheminformatics tools and application of computational methods for prediction of molecular fragments in protein binding sites. A summary of the conclusions from this research is presented below.

6.1 Unsatisfied donors/acceptors in protein-ligand complexes

An analysis of a set of 187 protein-ligand complexes showed that the total percentage of protein and ligand donor atoms that are inaccessible to water and are not explicitly hydrogen bonded is 3.23%. This percentage was found to be consistent with the earlier observations made for protein interiors[19]. Weak interactions, particularly CH-O and NH-π interactions were frequently associated with unsatisfied donors and acceptors. The fraction of unsatisfied donors and acceptors was used as a metric to discriminate between good and bad poses resulting from docking of ligand into their receptor structures in the same dataset. The results however did not show a direct relationship between RMSD of the top-scoring pose and the fraction of unsatisfied ligand donors and acceptors. It was therefore not possible to identify correct poses merely by a simple of count of missing hydrogen bonds. Recent studies with more sophisticated treatments, such as HYDE scoring function, where burial of polar groups lacking hydrogen bonds with ideal geometry is penalized, showed that only about a third of test cases showed better ranking of native-like poses than that which is achieved by another standard scoring scheme. The results from current analysis support the intuitively valid requirement of complete satisfaction of hydrogen bonding groups at buried protein-ligand interfaces

however fraction of unsatisfied donors and acceptors alone was not a useful metric for improvement in the ranking.

6.2 Weak aromatic interactions

As the importance of weak interactions in protein-ligand complexes is becoming more well-established, a survey of different types of ligand aromatic rings involved in interactions with CH and XH (X = N, O, S) groups in protein binding sites was conducted using IsoStar databse. The results of the survey indicate slight but meaningful differences in interaction preferences for different ring types. As expected, the presence of an electron-donating substituent on a phenyl ring was accompanied by higher CH/XH contact density above the ring plane. This geometry has been shown to be energetically more favourable and could include one of two isoenergetic configurations, T-shaped or parallel stacking geometry. An interesting observation was that this preference was slightly more obvious for CH groups than for XH groups. Similar observations were made in a recent survey where CH groups flanked by heteroatoms (N, O) showed higher above-ring preference than XH groups[10]. The modulation CH-π interactions by introducing electron donating substituents to an interacting ring moiety led to an increase in binding affinity of LCK kinase inhibitors[161]. Our database analysis therefore is therefore consistent with the idea that aromatic interactions have characteristic geometric patterns that probably have at least a 'supportive' role in ligand binding affinity in some cases.

6.3 Fragment docking and scoring with MM-GB/SA

Probing protein binding sites for energetically favourable positions for chemical moieties has been an elusive goal in computer-aided ligand design. Although pioneering methods date from before the advent of fragment-based lead discovery, there has been a renewed interest in developing robust methods for solvent mapping of protein binding sites or fragment docking and scoring due to successful FBLD campaigns against a variety of targets. The primary challenges however remain for scoring functions related to accurate treatment of solvent effects and target flexibility[67]. It was noticed from the application of MCSS on experimental solvent mapping datasets that a part of the problem in MCSS ranking was related to in overestimation of electrostatic interactions. Very crude approximation of solvent such as distance-dependent dielectric constant often leads to false positive energy minima. However, introduction of physically more accurate approaches such as implicit solvent models with continuum electrostatics led to improvement in scoring. This was noticed from the success rate (up to 75%) of MCSS (with MM-GB/SA) on a set of 12 protein-fragment complexes. Additionally, this scoring approach was also observed to benefit from the use of multiple structures in the case of receptors with multiple binding modes. Finally, good quality predictions were obtained for highly conserved water positions in the example of HSP90. Further improvement in parameterization of metal-ligand interactions, accurate treatment of water-mediated interactions and entropy estimation in the case of ligands belonging to different classes would enhance the applicability of MCSS followed by MM-GB/SA scoring.

6.4 Future Work

The results from this work have opened further opportunities in using MCSS as a robust method for mapping protein binding sites for fragment binding. The evaluation of MCSS minima with an implicit solvent based scoring function is however crucial. Further developments in MM-PB/ or MM-GB/SA scoring should enhance the applicability of the protocol of fragment docking and scoring described in this study.

The survey of unsatisfied hydrogen bond donors and acceptors done in this study further supports the intuitive expectation that a lost hydrogen bond should have a largely unfavourable effect on protein-ligand binding. The inclusion of such aspects of molecular recognition in docking and scoring should be a rewarding exercise. Although simple counts of the fractions of missing hydrogen bonds are not discriminatory enough, probably a more sophisticated approach would improve the ranking of candidate poses. This is supported by some recent developments in new scoring functions such as HYDE.

Finally, with the ongoing investigations and debate on the role of weak interactions in protein-ligand complexes, the survey in this study reflects to some extent that changes in interaction preferences of weak interactions follow intuitive chemical logic. The idea that at best a supportive role of weak interactions can be exploited in ligand design is becoming well-established and this study further supports this. Detailed theoretical and experimental investigation of weak interactions is, therefore, an interesting opportunity.

Bibliography

1. Dunn, M. F., Protein-Ligand Interactions: General Description. In *Encyclopedia of Life Sciences*, John Wiley & Sons, Ltd.: Chichester, 2010.

2. Fisher, H. F., Protein–Ligand Interactions:Molecular Basis. In *Encyclopedia of Life Sciences*, John Wiley & Sons, Ltd.: Chichester, 2001.

3. Perozzo, R.; Folkers, G.; Scapozza, L., Thermodynamics of protein-ligand interactions: History, presence, and future aspects. *Journal of Receptors and Signal Transduction* **2004,** 24, (1-2), 1-52.

4. Freire, E., Do enthalpy and entropy distinguish first in class from best in class? *Drug Discovery Today* **2008,** 13, (19-20), 869-874.

5. Gohlke, H.; Hendlich, M.; Klebe, G., Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspectives in Drug Discovery and Design* **2000,** 20, (1), 115-144.

6. Gohlke, H.; Klebe, G., Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie-International Edition* **2002,** 41, (15), 2645-2676.

7. Rich, R. L. R. L.; Myszka, D. G. D. G., Advances in surface plasmon resonance biosensor analysis. *Current Opinion in Biotechnology* **2000,** 11, 54-61.

8. Ruben, A. J.; Kiso, Y.; Freire, E., Overcoming roadblocks in lead optimization: A thermodynamic perspective. *Chemical Biology & Drug Design* **2006,** 67, (1), 2-4.

9. Ladbury, J. E., Calorimetry as a tool for understanding biomolecular interactions and an aid to drug design. *Biochemical Society Transactions* 38, 888-893.

10. Bissantz, C.; Kuhn, B.; Stahl, M., A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry* **2010,** 53, (14), 5061-5084.

11. Bohm, H. J., Prediction of Non-bonded Interactions in Drug Design. In *Protein-ligand Interactions: From Molecular Recognition to Drug Design*, Bohm, H. J.; Schneider, G., Eds. WILLEY-VCH Verlag GmbH & Co. KGaA: Weinham, 2003.

12. Hubbard, R. E.; Haider, M. K., Hydrogen Bonds in Proteins: Role and Strength. In *Encyclopedia of Life Sciences (ELS)*, John Wiley & Sons, Ltd: 2010.

13. Williams, M. A.; Ladbury, J. E., Hydrogen Bonds in Protein-Ligand Complexes. In *Protein-Ligand Interactions: From Molecular Recognition to Drug Design*, Bohm, H. J.; Schneider, G., Eds. WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, 2003.

14. Bruno, I. J.; Cole, J. C.; Lommerse, J. P. M.; Rowland, R. S.; Taylor, R.; Verdonk, M. L., IsoStar: A library of information about nonbonded interactions. *Journal of Computer-Aided Molecular Design* **1997,** 11, (6), 525-537.

15. Boer, D. R.; Kroon, J.; Cole, J. C.; Smith, B.; Verdonk, M. L., SuperStar: Comparison of CSD and PDB-based interaction fields as a basis for the prediction of protein-ligand interactions. *Journal of Molecular Biology* **2001,** 312, (1), 275-287.

16. Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G., Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *Journal of Molecular Biology* **2003,** 326, (2), 607-620.

17. Fersht, A. R.; Shi, J. P.; Knilljones, J.; Lowe, D. M.; Wilkinson, A. J.; Blow, D. M.; Brick, P.; Carter, P.; Waye, M. M. Y.; Winter, G., Hydrogen-bonding and biological specificity analyzed by protein engineering. *Nature* **1985,** 314, (6008), 235-238.

18. Shirley, B. A.; Stanssens, P.; Hahn, U.; Pace, C. N., Contribution of hydrogen-bonding to the conformational stability of ribonuclease-t1. *Biochemistry* **1992,** 31, (3), 725-732.

19. McDonald, I. K.; Thornton, J. M., Satisfying hydrogen-bonding potential in proteins. *Journal of Molecular Biology* **1994,** 238, (5), 777-793.

20. Richards, F. M., Areas, Volumes, Packing, and Protein Structure. *Annual Review of Biophysics and Bioengineering* **1977,** 6, 151-176.

21. Sharp, K. A.; Nicholls, A.; Friedman, R.; Honig, B., Extracting Hydrophobic Free Energies from Experimental Data - Relationship to Protein Folding and Theoretical Models. *Biochemistry* **1991,** 30, (40), 9686-9697.

22. Peters, J.-U.; Weber, S.; Kritter, S.; Weiss, P.; Wallier, A.; Boehringer, M.; Hennig, M.; Kuhn, B.; Loeffler, B.-M., Aminomethylpyrimidines as novel DPP-IV inhibitors: A 105-fold activity increase by optimization of aromatic substituents. *Bioorganic & Medicinal Chemistry Letters* **2004,** 14, (6), 1491-1493.

23. Meyer, E. A.; Castellano, R. K.; Diederich, F., Interactions with aromatic rings in chemical and biological recognition. *Angewandte Chemie-International Edition* **2003,** 42, (11), 1210-1250.

24. McGaughey, G. B.; Gagne, M.; Rappe, A. K., pi-stacking interactions - Alive and well in proteins. *Journal of Biological Chemistry* **1998,** 273, (25), 15458-15463.

25. Desiraju, G. R., CH...O and other weak hydrogen bonds. From crystal engineering to virtual screening. *Chemical Communications* **2005**, (24), 2995-3001.

26. Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K., The Magnitude of the CH/pi Interaction between Benzene and Some Model Hydrocarbons. *Journal of the American Chemical Society* **2000,** 122, (15), 3746-3753.

27. Wissner, A.; Berger, D. M.; Boschelli, D. H.; Floyd, M. B.; Greenberger, L. M.; Gruber, B. C.; Johnson, B. D.; Mamuya, N.; Nilakantan, R.; Reich, M. F.; Shen, R.; Tsou, H. R.; Upeslacis, E.; Wang, Y. F.; Wu, B. Q.; Ye, F.; Zhang, N., 4-Anilino-6,7-dialkoxyquinolime-3-carbonitrile inhibitors of epidermal growth factor receptor kinase and their bioisosteric relationship to the 4-anilino-6,7-dialkoxyquinazoline inhibitors. *Journal of Medicinal Chemistry* **2000,** 43, (17), 3244-3256.

28. Liu, C. J.; Wrobleski, S. T.; Lin, J.; Ahmed, G.; Metzger, A.; Wityak, J.; Gillooly, K. M.; Shuster, D. J.; McIntyre, K. W.; Pitt, S.; Shen, D. R.; Zhang, R. F.; Zhang, H. J.; Doweyko, A. M.; Diller, D.; Henderson, I.; Barrish, J. C.; Dodd, J. H.; Schieven, G. L.; Leftheris, K., 5-cyanopyrimidine derivatives as a novel class of potent, selective, and orally active inhibitors of p38 alpha MAP kinase. *Journal of Medicinal Chemistry* **2005,** 48, (20), 6261-6270.

29. Campiani, G.; Kozikowski, A. P.; Wang, S. M.; Ming, L.; Nacci, V.; Saxena, A.; Doctor, B. P., Synthesis and anticholinesterase activity of huperzine A analogues containing phenol and catechol replacements for the pyridone ring. *Bioorganic & Medicinal Chemistry Letters* **1998,** 8, (11), 1413-1418.

30. Scatena, L. F.; Brown, M. G.; Richmond, G. L., Water at hydrophobic surfaces: Weak hydrogen bonding and strong orientation effects. *Science* **2001,** 292, (5518), 908-912.

31. Finney, J. L.; Soper, A. K., Solvent Structure and Perturbations in Solutions of Chemical and Biological Importance. *Chemical Society Reviews* **1994,** 23, (1), 1-10.

32. Piatnitski, E. L.; Flowers Ii, R. A.; Deshayes, K., Highly Organized Spherical Hosts That Bind Organic Guests in Aqueous Solution with Micromolar Affinity: Microcalorimetry Studies. *Chemistry – A European Journal* **2000,** 6, (6), 999-1006.

33. Homans, S. W., Water, water everywhere - except where it matters? *Drug Discovery Today* **2007,** 12, (13-14), 534-539.

34. Talhout, R.; Villa, A.; Mark, A. E.; Engberts, J., 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. *Journal of the American Chemical Society* **2003,** 125, (35), 10570-10579.

35. Mobley, D. L.; Dill, K. A., Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure* **2009,** 17, (4), 489-498.

36. Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A., Target Flexibility: An Emerging Consideration in Drug Discovery and Design. *Journal of Medicinal Chemistry* **2008,** 51, (20), 6237-6255.

37. Ali, M. M. U.; Roe, S. M.; Vaughan, C. K.; Meyer, P.; Panaretou, B.; Piper, P. W.; Prodromou, C.; Pearl, L. H., Crystal structure of an Hsp90â€“nucleotideâ€“p23/Sba1 closed chaperone complex. *Nature* **2006,** 440, (7087), 1013-1017.

38. Colombo, G.; Morra, G.; Meli, M.; Verkhivker, G., Understanding ligand-based modulation of the Hsp90 molecular chaperone dynamics at atomic resolution. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** 105, (23), 7976-7981.

39. Gutteridge, A.; Thornton, J., Conformational changes observed in enzyme crystal structures upon substrate binding. *Journal of Molecular Biology* **2005,** 346, (1), 21-28.

40. Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M., Side-chain flexibility in proteins upon ligand binding. *Proteins-Structure Function and Genetics* **2000,** 39, (3), 261-268.

41. Shiau, A. K.; Harris, S. F.; Southworth, D. R.; Agard, D. A., Structural Analysis of E. coli hsp90 Reveals Dramatic Nucleotide-Dependent Conformational Rearrangements. *Cell* **2006,** 127, (2), 329-340.

42. Richter, K.; Buchner, J., hsp90: Twist and Fold. *Cell* **2006,** 127, (2), 251-253.

43. Huth, J. R.; Park, C.; Petros, A. M.; Kunzer, A. R.; Wendt, M. D.; Wang, X. L.; Lynch, C. L.; Mack, J. C.; Swift, K. M.; Judge, R. A.; Chen, J.; Richardson, P. L.; Jin, S.; Tahir, S. K.; Matayoshi, E. D.; Dorwin, S. A.; Ladror, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; Fesik, S. W.; Elmore, S. W.; Hajduk, P. J., Discovery and design of novel HSP90 inhibitors using multiple fragment-based design strategies. *Chemical Biology & Drug Design* **2007,** 70, 1-12.

44. Brough, P. A.; Barril, X.; Borgognoni, J.; Chene, P.; Davies, N. G. M.; Davis, B.; Drysdale, M. J.; Dymock, B.; Eccles, S. A.; Garcia-Echeverria, C.; Fromont, C.; Hayes, A.; Hubbard, R. E.; Jordan, A. M.; Jensen, M. R.; Massey, A.; Merrett, A.; Padfield, A.; Parsons, R.; Radimerski, T.; Raynaud, F. I.; Robertson, A.; Roughley, S. D.; Schoepfer, J.; Simmonite, H.; Sharp, S. Y.; Surgenor, A.; Valenti, M.; Walls, S.; Webb, P.; Wood, M.; Workman, P.; Wright, L., Combining Hit Identification Strategies: Fragment-

Based and in Silico Approaches to Orally Active 2-Aminothieno 2,3-d pyrimidine Inhibitors of the Hsp90 Molecular Chaperone. *Journal of Medicinal Chemistry* **2009,** 52, (15), 4794-4809.

45. Pertinhez, T.; Sforça, M.; Alves, A.; Ramos, C.; Ho, P.; Tendler, M.; Zanchin, N.; Spisni, A., Letter to the Editor: ¹H, ¹⁵N and ¹³C resonance assignments of the apo Sm14-M20(C62V) protein, a mutant of <i>Schistosoma mansoni</i> Sm14. *Journal of Biomolecular NMR* **2004,** 29, (4), 553-554.

46. Fanelli, F.; De Benedetti, P. G., Computational Modeling Approaches to Structure-Function Analysis of G Protein-Coupled Receptors. *Chemical Reviews* **2005,** 105, (9), 3297-3351.

47. Massova, I.; Kollman, P. A., Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspectives in Drug Discovery and Design* **2000,** 18, 113-135.

48. Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: The Biomolecular Simulation Program. *Journal of Computational Chemistry* **2009,** 30, (10), 1545-1614.

49. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *Journal of the American Chemical Society* **1995,** 117, (19), 5179-5197.

50. Feig, M.; Brooks, C. L., Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Current Opinion in Structural Biology* **2004,** 14, (2), 217-224.

51. Bohm, H. J.; Stahl, M., Rapid empirical scoring functions in virtual screening applications. *Medicinal Chemistry Research* **1999,** 9, (7-8), 445-462.

52. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions .1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design* **1997,** 11, (5), 425-445.

53. Schulz-Gasch, T.; Stahl, M., Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discovery Today: Technologies* **2004,** 1, (3), 231-239.

54. Fenu, L. A.; Lewis, R. A.; Good, A. C.; Bodkin, M.; Essex, J. W., SCORING FUNCTIONS: From Free-energies of Binding to Enrichment in Virtual screening. In *Structure-based Drug Discovery*, Jhoti, H.; Leach, A., Eds. Springer: 2007; pp 223-245.

55. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M., BLEEP—potential of mean force describing protein–ligand interactions: I. Generating potential. *Journal of Computational Chemistry* **1999,** 20, (11), 1165-1176.

56. Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *Journal of Medicinal Chemistry* **1999,** 42, (5), 791-804.

57. MacKerell Jr, A. D., Atomistic Models and Force-fields. In *Computational Biochemistry and Biophysics*, Becker et al, O. M., Ed. Marcel Dekker, Inc.: New York, 2001.

58. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., Assessing scoring functions for protein-ligand interactions. *Journal of Medicinal Chemistry* **2004,** 47, (12), 3032-3047.

59. Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J., Recent developments in fragment-based drug discovery. *Journal of Medicinal Chemistry* **2008,** 51, (13), 3661-3680.

60. Fischer, M.; Hubbard, R. E., Fragment-based ligand discovery. *Molecular Interventions* **2009,** 9, (1), 22-30.

61. Hajduk, P. J.; Greer, J., A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews Drug Discovery* **2007,** 6, (3), 211-219.

62. Hajduk, P. J., SAR by NMR: Putting the pieces together. *Molecular Interventions* **2006,** 6, (5), 266-+.

63. Erlanson, D. A.; McDowell, R. S.; O'Brien, T., Fragment-based drug discovery. *Journal of Medicinal Chemistry* **2004,** 47, (14), 3463-3482.

64. Zartler, E. R.; Shapiro, M. J., Fragonomics: fragment-based drug discovery. *Current Opinion in Chemical Biology* **2005,** 9, (4), 366-370.

65. de Kloe, G. E.; Bailey, D.; Leurs, R.; de Esch, I. J. P., Transforming fragments into candidates: small becomes big in medicinal chemistry. *Drug Discovery Today* **2009,** 14, (13-14), 630-646.

66. Chen, Y.; Shoichet, B. K., Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nature Chemical Biology* **2009,** 5, (5), 358-364.

67. Law, R.; Barker, O.; Barker, J. J.; Hesterkamp, T.; Godemann, R.; Andersen, O.; Fryatt, T.; Courtney, S.; Hallett, D.; Whittaker, M., The multiple roles of computational chemistry in fragment-based drug design. *Journal of Computer-Aided Molecular Design* **2009,** 23, (8), 459-473.

68. Goodford, P. J., A Computational-Procedure For Determining Energetically Favorable Binding-Sites On Biologically Important Macromolecules. *Journal of Medicinal Chemistry* **1985,** 28, (7), 849-857.

69. Miranker, A.; Karplus, M., Functionality Maps Of Binding-Sites - A Multiple Copy Simultaneous Search Method. *Proteins-Structure Function and Genetics* **1991,** 11, (1), 29-34.

70. Bitetti-Putzer, R.; Joseph-McCarthy, D.; Hogle, J. M.; Karplus, M., Functional group placement in protein binding sites: a comparison of GRID and MCSS. *Journal of Computer-Aided Molecular Design* **2001,** 15, (10), 935-960.

71. Cross, S.; Cruciani, G., Molecular fields in drug discovery: getting old or reaching maturity? *Drug Discovery Today* **2010,** 15, (1-2), 23-32.

72. Caflisch, A.; Miranker, A.; Karplus, M., Multiple copy simultaneous search and construction of ligands in binding sites - application to inhibitors of HIV-1 aspartic proteinase. *Journal of Medicinal Chemistry* **1993,** 36, (15), 2142-2167.

73. Eisen, M. B.; Wiley, D. C.; Karplus, M.; Hubbard, R. E., HOOK: A Program for Finding Novel Molecular Architectures That Satisfy the Chemical and Steric Requirements of a Macromolecule Binding Site. *Proteins: Structure, Function and Genetics* **1994,** 19, (3), 199-221.

74. Caflisch, A., Computational combinatorial ligand design: Application to human alpha-thrombin. *Journal of Computer-Aided Molecular Design* **1996,** 10, (5), 372-396.

75. Joseph-McCarthy, D.; Hogle, J. M.; Karplus, M., Use of the multiple copy simultaneous search (MCSS) method to design a new class of picornavirus capsid binding drugs. *Proteins-Structure Function and Bioinformatics* **1997,** 29, (1), 32-58.

76. Schubert, C. R.; Stultz, C. M., The multi-copy simultaneous search methodology: a fundamental tool for structure-based drug design. *Journal of Computer-Aided Molecular Design* **2009,** 23, (8), 475-489.

77. Allen, K. N.; Bellamacina, C. R.; Ding, X. C.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D., An experimental approach to mapping the binding surfaces of crystalline proteins. *Journal of Physical Chemistry* **1996,** 100, (7), 2605-2611.

78. Mattos, C.; Ringe, D., Locating and characterizing binding sites on proteins. *Nature Biotechnology* **1996,** 14, (5), 595-599.

79. Joseph-McCarthy, D.; Fedorov, A. A.; Almo, S. C., Comparison of experimental and computational functional group mapping of an RNase A structure: Implications for computer-aided drug design. *Protein Engineering* **1996,** 9, (9), 773-780.

80. English, A. C.; Done, S. H.; Caves, L. S. D.; Groom, C. R.; Hubbard, R. E., Locating interaction sites on proteins: The crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins-Structure Function and Genetics* **1999,** 37, (4), 628-640.

81. English, A. C.; Groom, C. R.; Hubbard, R. E., Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Engineering* **2001,** 14, (1), 47-59.

82. Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D., Multiple solvent crystal structures: Probing binding sites, plasticity and hydration. *Journal of Molecular Biology* **2006,** 357, (5), 1471-1482.

83. Guimaraes, C. R. W.; Cardozo, M., MM-GB/SA rescoring of docking poses in structure-based lead optimization. *Journal of Chemical Information and Modeling* **2008,** 48, (5), 958-970.

84. Michel, J.; Verdonk, M. L.; Essex, J. W., Protein-ligand binding affinity predictions by implicit solvent simulations: A tool for lead optimization? *Journal of Medicinal Chemistry* **2006,** 49, (25), 7427-7439.

85. Thompson, D. C.; Humblet, C.; Joseph-McCarthy, D., Investigation of MM-PBSA rescoring of docking poses. *Journal of Chemical Information and Modeling* **2008,** 48, (5), 1081-1091.

86. Dennis, S.; Kortvelyesi, T.; Vajda, S., Computational mapping identifies the binding sites of organic solvents on proteins. *Proceedings of the National Academy of Sciences* **2002,** 99, (7), 4290-4295.

87. Silberstein, M.; Dennis, S.; Brown, L.; Kortvelyesi, T.; Clodfelter, K.; Vajda, S., Identification of substrate binding sites in enzymes by computational solvent mapping. *Journal of Molecular Biology* **2003,** 332, (5), 1095-1113.

88. Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S., Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009,** 25, (5), 621-627.

89. Landon, M. R.; Lieberman, R. L.; Hoang, Q. Q.; Ju, S. L.; Caaveiro, J. M. M.; Orwig, S. D.; Kozakov, D.; Brenke, R.; Chuang, G. Y.; Beglov, D.; Vajda, S.; Petsko, G. A.; Ringe, D., Detection of ligand binding hot spots on protein surfaces via fragment-based methods: application to DJ-1 and glucocerebrosidase. *Journal of Computer-Aided Molecular Design* **2009,** 23, (8), 491-500.

90. Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A., Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins-Structure Function and Genetics* **1999,** 37, (1), 88-105.

91. Majeux, N.; Scarsi, M.; Caflisch, A., Efficient electrostatic solvation model for protein-fragment docking. *Proteins-Structure Function and Genetics* **2001,** 42, (2), 256-268.

92. Schaefer, M.; Karplus, M., A comprehensive analytical treatment of continuum electrostatics. *Journal of Physical Chemistry* **1996,** 100, (5), 1578-1599.

93. Budin, N.; Majeux, N.; Caflisch, A., Fragment-based flexible ligand docking by evolutionary optimization. *Biological Chemistry* **2001,** 382, (9), 1365-1372.

94. Huang, D.; Caflisch, A., Efficient evaluation of binding free energy using continuum electrostatics solvation. *Journal of Medicinal Chemistry* **2004,** 47, (23), 5791-5797.

95. Huang, D.; Caflisch, A., Library screening by fragment-based docking. *Journal of Molecular Recognition* **2009,** 23, (2), 183Â -Â 193.

96. Imai, T.; Oda, K.; Kovalenko, A.; Hirata, F.; Kidera, A., Ligand Mapping on Protein Surfaces by the 3D-RISM Theory: Toward Computational Fragment-Based Drug Design. *Journal of the American Chemical Society* **2009,** 131, (34), 12430-12440.

97. Beglov, D.; Roux, B., An integral equation to describe the solvation of polar molecules in liquid water. *Journal of Physical Chemistry B* **1997,** 101, (39), 7821-7826.

98. Kovalenko, A.; Hirata, F., Three-dimensional density profiles of water in contact with a solute of arbitrary shape: A RISM approach. *Chemical Physics Letters* **1998,** 290, (1-3), 237-244.

99. Leach, A. R., *Molecular Modelling: Principles and Applications*. 2nd ed.; Pearson Education EMA: 2001.

100. Onufriev, A., Continuum Electrostatics Solvent Modeling with the Generalized B orn Model. In *Modeling Solvent Environments Applications to Simulations of Biomolecules*, Feig, M., Ed. WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, 2010.

101. Feig, M.; Chocholousova, J.; Tanizaki, S., Simulating nucleic acids: Towards implicit and hybrid solvation models. *Biophysical Journal* **2005,** 88, (1), 513A-513A.

102. Baker, N. A., Improving implicit solvent simulations: a Poisson-centric view. *Current Opinion in Structural Biology* **2005,** 15, (2), 137-143.

103. Dill, K. A.; Bromberg, S., *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology*. Taylor & Francis Group: 2002.

104. Onufriev, A.; Case, D. A.; Bashford, D., Effective Born radii in the generalized Born approximation: The importance of being perfect. *Journal of Computational Chemistry* **2002,** 23, (14), 1297-1304.

105. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., SEMIANALYTICAL TREATMENT OF SOLVATION FOR MOLECULAR MECHANICS AND DYNAMICS. *Journal of the American Chemical Society* **1990,** 112, (16), 6127-6129.

106. Grycuk, T., Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation. *Journal of Chemical Physics* **2003,** 119, (9), 4817-4826.

107. Lee, M. S.; Feig, M.; Salsbury Jr, F. R.; Brooks Iii, C. L., New analytical approximation to the standard molecular volume definition and its application to generalized Born calculations. *Journal of Computational Chemistry* **2003,** 24, (11), 1348Â -Â 1356.

108. Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K., On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy. *Journal of the American Chemical Society* **2003,** 125, (31), 9523-9530.

109. Baker, E. N.; Hubbard, R. E., Hydrogen-bonding in globular-proteins. *Progress in Biophysics & Molecular Biology* **1984,** 44, (2), 97-179.

110. Pauling, L.; Corey, R. B., The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. *Proceedings of the National Academy of Sciences of the United States of America* **1951,** 37, (5), 251-256.

111. Pauling, L.; Corey, R. B.; Branson, H. R., The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proceedings of the National Academy of Sciences of the United States of America* **1951,** 37, (4), 205-211.

112. Savage, H. J.; Elliott, C. J.; Freeman, C. M.; Finney, J. L., Lost hydrogen-bonds and buried surface-area - rationalizing stability in globular-proteins. *Journal of the Chemical Society-Faraday Transactions* **1993,** 89, (15), 2609-2617.

113. Joh, N. H.; Min, A.; Faham, S.; Whitelegge, J. P.; Yang, D.; Woods, V. L.; Bowie, J. U., Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins. *Nature* **2008,** 453, (7199), 1266-U73.

114. Fleming, P. J.; Rose, G. D., Do all backbone polar groups in proteins form hydrogen bonds? *Protein Science* **2005,** 14, (7), 1911-1917.

115. Barratt, E.; Bronowska, A.; Vondrásek, J.; Cerný, J.; Bingham, R.; Phillips, S.; Homans, S. W., Thermodynamic Penalty Arising from Burial of a Ligand Polar Group Within a Hydrophobic Pocket of a Protein Receptor. *Journal of Molecular Biology* **2006,** 362, (5), 994-1003.

116. Giordanetto, F.; Cotesta, S.; Catana, C.; Trosset, J. Y.; Vulpetti, A.; Stouten, P. F. W.; Kroemer, R. T., Novel scoring functions comprising QXP, SASA, and protein side-chain entropy terms. *Journal of Chemical Information and Computer Sciences* **2004,** 44, (3), 882-893.

117. Reulecke, I.; Lange, G.; Albrecht, J.; Klein, R.; Rarey, M., Towards an integrated description of hydrogen bonding and dehydration: Decreasing false positives in virtual screening with the HYDE scoring function. *Chemmedchem* **2008,** 3, (6), 885-897.

118. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology* **1996,** 261, (3), 470-489.

119. Kellogg, G. E.; Burnett, J. C.; Abraham, D. J., Very empirical treatment of solvation and entropy: a force field derived from Log P-o/w. *Journal of Computer-Aided Molecular Design* **2001,** 15, (4), 381-393.

120. Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R., A new test set for validating predictions of protein-ligand interaction. *Proteins-Structure Function and Genetics* **2002,** 49, (4), 457-471.

121. Sanner, M. F.; Coon, I. S. *MolKit Package, MGL Tools*, 2002.

122. McDonald, I.; Thornton, J. M., Atlas of Side-Chain and Main-Chain Hydrogen Bonding. In WWW Edition December 1994 ed.; 1994.

123. Hubbard, S. J.; Thornton, J. M. *NACCESS*, Department of Biochemistry and Molecular Biology, University College London: 1993.

124. Lee, B.; Richards, F. M., Interpretation of protein structures - estimation of static accessibility. *Journal of Molecular Biology* **1971,** 55, (3), 379-&.

125. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C., Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology* **1999,** 285, (4), 1735-1747.

126. Parthasarathy, S.; Murthy, M. R. N., Analysis of temperature factor distribution in high-resolution protein structures. *Protein Science* **1997,** 6, (12), 2561-2567.

127. Morley, S. D.; Afshar, M., Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock (R). *Journal of Computer-Aided Molecular Design* **2004,** 18, (3), 189-208.

128. Community, O. B. *The Open Babel Chemical File Format Conversion Package*, http://openbabel.sourceforge.net/: 2008.

129. Parthasarathy, S.; Murthy, M. R. N., Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Engineering* **2000,** 13, (1), 9-13.

130. Scholtz, J. M.; Marqusee, S.; Baldwin, R. L.; York, E. J.; Stewart, J. M.; Santoro, M.; Bolen, D. W., Calorimetric determination of the enthalpy change for the alpha-helix to coil transition of an alanine peptide in water. *PNAS* **1991,** 88, (7), 2854-2858.

131. Makhatadze, G. I.; Privalov, P. L., Contribution of hydration to protein-folding thermodynamics .1. The enthalpy of hydration. *Journal of Molecular Biology* **1993,** 232, (2), 639-659.

132. Myers, J. K.; Pace, C. N., Hydrogen bonding stabilizes globular proteins. *Biophysical Journal* **1996,** 71, (4), 2033-2039.

133. Shortle, D., Propensities, probabilities, and the Boltzmann hypothesis. *Protein Science* **2003,** 12, (6), 1298-1302.

134. Tóth, G.; Bowers, S. G.; Truong, A. P.; Probst, G., The Role and Significance of Unconventional Hydrogen Bonds in Small Molecule Recognition by Biological

Receptors of Pharmaceutical Relevance. *Current Pharmaceutical Design* **2007,** 13, 3476-3493.

135. Klaholz, B. P.; Moras, D., C-H···O Hydrogen Bonds in the Nuclear Receptor RAR[gamma]--a Potential Tool for Drug Selectivity. *Structure* **2002,** 10, (9), 1197-1204.

136. Wallace, A. C.; Laskowski, R. A.; Thornton, J. M., LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. *Prot. Eng.* **1995,** 8, 127-134.

137. Steiner, T.; Koellner, G., Hydrogen bonds with pi-acceptors in proteins: Frequencies and role in stabilizing local 3D structures. *Journal of Molecular Biology* **2001,** 305, (3), 535-557.

138. Desiraju, G. R.; Steiner, T., *Weak Hydrogen Bond in Structural Chemistry and Biology*. Oxforf University Press: 1999.

139. Perutz, M. F., The Role of Aromatic Rings as Hydrogen-Bond Acceptors in Molecular Recognition. *Philosophical Transactions of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences* **1993,** 345, (1674), 105-112.

140. Wlodawer, A.; Walter, J.; Huber, R.; Sjolin, L., Structure of Bovine Pancreatic Trypsin-Inhibitor - Results of Joint Neutron and X-Ray Refinement of Crystal Form-II. *Journal of Molecular Biology* **1984,** 180, (2), 301-329.

141. Chakkaravarthi, S.; Babu, M. M.; Gromiha, M. M.; Jayaraman, G.; Sethumadhavan, R., Exploring the environmental preference of weak interactions in (alpha/beta)(8) barrel proteins. *Proteins-Structure Function and Bioinformatics* **2006,** 65, (1), 75-86.

142. Sarkhel, S.; Desiraju, G. R., NH...O, OH...O, and CH...O hydrogen bonds in protein-ligand complexes: Strong and weak interactions in molecular recognition. *Proteins-Structure Function and Genetics* **2004,** 54, (2), 247-259.

143. Madan Babu, M.; Kumar Singh, S.; Balaram, P., A C-H...O Hydrogen Bond Stabilized Polypeptide Chain Reversal Motif at the C Terminus of Helices in Proteins. *Journal of Molecular Biology* **2002,** 322, (4), 871-880.

144. Fabiola, G. F.; Krishnaswamy, S.; Nagarajan, V.; Pattabhi, V., C-H...O Hydrogen Bonds in β-sheets. *Acta Crystallographica Section D* **1997,** 53, (3), 316-320.

145. Cappelli, A.; Giorgi, G.; Anzini, M.; Vomero, S.; Ristori, S.; Rossi, C.; Donati, A., Characterization of Persistent Intramolecular C-H...X(N,O) Bonds in Solid State and Solution. *Chemistry – A European Journal* **2004,** 10, (13), 3177-3183.

146. Thallapally, P. K.; Katz, A. K.; Carrell, H. L.; Desiraju, G. R., Unusually long cooperative chain of seven hydrogen bonds. An alternative packing type for symmetrical phenols. *Chemical Communications* **2002**, (4), 344-345.

147. Hof, F.; Diederich, F., Medicinal chemistry in academia: molecular recognition with biological receptors. *Chemical Communications* **2004**, (5), 477-480.

148. Allen, F. H., The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B-Structural Science* **2002,** 58, 380-388.

149. Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z. K.; Gilliland, G.; Weissig, H.; Westbrook, J., The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology* **2000,** 7, 957-959.

150. Imai, Y. N.; Inoue, Y.; Nakanishi, I.; Kitaura, K., Cl-pi interactions in protein-ligand complexes. *Protein Science* **2008,** 17, (7), 1129-1137.

151. Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K., Origin of the attraction and directionality of the NH/pi interaction: Comparison with OH/pi and CH/pi interactions. *Journal of the American Chemical Society* **2000,** 122, (46), 11450-11458.

152. Courty, A.; Mons, M.; Dimicoli, I.; Piuzzi, F.; Gaigeot, M.-P.; Brenner, V.; de Pujo, P.; Millie, P., Quantum Effects in the Threshold Photoionization and Energetics of the Benzeneâˆ’H2O and Benzeneâˆ’D2O Complexes: Experiment and Simulation. *The Journal of Physical Chemistry A* **1998,** 102, (33), 6590-6600.

153. Mons, M.; Dimicoli, I.; Tardivel, B.; Piuzzi, F.; Brenner, V.; Millie, P., Energetics of a model NH-pi interaction: the gas phase benzene-NH3 complex. *Physical Chemistry Chemical Physics* **2002,** 4, (4), 571-576.

154. Chakrabarti, P.; Chakrabarti, S., CH...O hydrogen bond involving proline residues in alpha-helices. *Journal of Molecular Biology* **1998,** 284, (4), 867-873.

155. Derewendra, Z. S.; Lee, L.; Derewenda, U., The Occurrence of CH...O Hydrogen Bonds in Proteins. *Journal of Molecular Biology* **1995,** 252, (2), 248-262.

156. Duan, G.; Smith, V. H.; Weaver, D. F., Characterization of aromatic-amide(side- chain) interactions in proteins through systematic ab initio calculations and data mining analyses. *Journal of Physical Chemistry A* **2000,** 104, (19), 4521-4532.

157. Wang, S. M.; Liu, M.; Lewin, N. E.; Lorenzo, P. S.; Bhattacharrya, D.; Qiao, L. X.; Kozikowski, A. P.; Blumberg, P. M., Probing the binding of indolactam-V to protein

kinase C through site-directed mutagenesis and computational docking simulations. *Journal of Medicinal Chemistry* **1999,** 42, (18), 3436-3446.

158. Schoepfer, J.; Gay, B.; Caravatti, G.; Garcia-Echeverria, C.; Fretz, H.; Rahuel, J.; Furet, P., Structure-based design of peptidomimetic ligands of the Grb2-SH2 domain. *Bioorganic & Medicinal Chemistry Letters* **1998,** 8, (20), 2865-2870.

159. Pierce, A. C.; ter Haar, E.; Binch, H. M.; Kay, D. P.; Patel, S. R.; Li, P., CH...O and CH...N hydrogen bonds in ligand design: A novel quinazolin-4-ylthiazol-2-ylamine protein kinase inhibitor. *Journal of Medicinal Chemistry* **2005,** 48, (4), 1278-1281.

160. Pierce, A. C.; Sandretto, K. L.; Bemis, G. W., Kinase inhibitors and the case for CH...O hydrogen bonds in protein-ligand binding. *Proteins-Structure Function and Genetics* **2002,** 49, (4), 567-576.

161. Ozawa, T.; Tsuji, E.; Ozawa, M.; Handa, C.; Mukaiyama, H.; Nishimura, T.; Kobayashi, S.; Okazaki, K., The importance of CH/[pi] hydrogen bonds in rational drug design: An ab initio fragment molecular orbital study to leukocyte-specific protein tyrosine (LCK) kinase. *Bioorganic & Medicinal Chemistry* **2008,** 16, (24), 10311-10318.

162. Gosling, M. P.; Pugliesi, I.; Cockett, M. C. R., The role of the methyl group in stabilising the weak N-H[small pi] hydrogen bond in the 4-fluorotoluene-ammonia complex. *Physical Chemistry Chemical Physics* 12, (1), 132-142.

163. Whitesell, L.; Lindquist, S. L., HSP90 and the chaperoning of cancer. *Nature Reviews Cancer* **2005,** 5, (10), 761-772.

164. Selinsky, B. S.; Gupta, K.; Sharkey, C. T.; Loll, P. J., Structural analysis of NSAID binding by prostaglandin H-2 synthase: Time-dependent and time-independent inhibitors elicit identical enzyme conformations. *Biochemistry* **2001,** 40, (17), 5172-5180.

165. Hajduk, P. J.; Boyd, S.; Nettesheim, D.; Nienaber, V.; Severin, J.; Smith, R.; Davidson, D.; Rockway, T.; Fesik, S. W., Identification of novel inhibitors of urokinase via NMR-based screening. *Journal of Medicinal Chemistry* **2000,** 43, (21), 3862-3866.

166. Warnmark, A.; Treuter, E.; Gustafsson, J. A.; Hubbard, R. E.; Brzozowski, A. M.; Pike, A. C. W., Interaction of transcriptional intermediary factor 2 nuclear receptor box peptides with the coactivator binding site of estrogen receptor alpha. *Journal of Biological Chemistry* **2002,** 277, (24), 21862-21868.

167. Rasmussen, H. B.; Branner, S.; Wiberg, F. C.; Wagtmann, N., Crystal structure of human dipeptidyl peptidase IV/CD26 in complex with a substrate analog. *Nature Structural Biology* **2003,** 10, (1), 19-25.

168. Meyer, E. A.; Furler, M.; Diederich, F.; Brenk, R.; Klebe, G., Synthesis and In Vitro Evaluation of 2-Aminoquinazolin-4(3H)-one-Based Inhibitors for tRNA-Guanine Transglycosylase (TGT). *Helvetica Chimica Acta* **2004,** 87, (6), 1333-1356.

169. Hartshorn, M. J.; Murray, C. W.; Cleasby, A.; Frederickson, M.; Tickle, I. J.; Jhoti, H., Fragment-based lead discovery using X-ray crystallography. *Journal of Medicinal Chemistry* **2005,** 48, (2), 403-413.

170. Wu, Q.; Gee, C. L.; Lin, F.; Tyndall, J. D.; Martin, J. L.; Grunewald, G. L.; McLeish, M. J., Structural, mutagenic, and kinetic analysis of the binding of substrates and inhibitors of human phenylethanolamine N-methyltransferase. *Journal of Medicinal Chemistry* **2005,** 48, (23), 7243-7252.

171. Kallander, L. S.; Lu, Q.; Chen, W. F.; Tomaszek, T.; Yang, G.; Tew, D.; Meek, T. D.; Hofmann, G. A.; Schulz-Pritchard, C. K.; Smith, W. W.; Janson, C. A.; Ryan, M. D.; Zhang, G. F.; Johanson, K. O.; Kirkpatrick, R. B.; Ho, T. F.; Fisher, P. W.; Mattern, M. R.; Johnson, R. K.; Hansbury, M. J.; Winkler, J. D.; Ward, K. W.; Veber, D. F.; Thompson, S. K., 4-Aryl-1,2,3-triazole: A novel template for a reversible methionine aminopeptidase 2 inhibitor, optimized to inhibit angiogenesis in vivo. *Journal of Medicinal Chemistry* **2005,** 48, (18), 5644-5647.

172. Howard, N.; Abell, C.; Blakemore, W.; Chessari, G.; Congreve, M.; Howard, S.; Jhoti, H.; Murray, C. W.; Seavers, L. C. A.; van Montfort, R. L. M., Application of fragment screening and fragment linking to the discovery of novel thrombin inhibitors. *Journal of Medicinal Chemistry* **2006,** 49, (4), 1346-1355.

173. Murray, C. W.; Callaghan, O.; Chessari, G.; Cleasby, A.; Congreve, M.; Frederickson, M.; Hartshorn, M. J.; McMenamin, R.; Patel, S.; Wallis, N., Application of fragment screening by X-ray crystallography to beta-secretase. *Journal of Medicinal Chemistry* **2007,** 50, (6), 1116-1123.

174. Dutta, R.; Inouye, M., GHKL, an emergent ATPase/kinase superfamily. *Trends in Biochemical Sciences* **2000,** 25, (1), 24-28.

175. Solit, D. B.; Chiosis, G., Development and application of Hsp90 inhibitors. *Drug Discovery Today* **2008,** 13, (1-2), 38-43.

176. Soldano, K. L.; Jivan, A.; Nicchitta, C. V.; Gewirth, D. T., Structure of the N-terminal domain of GRP94 - Basis for ligand specificity and regulation. *Journal of Biological Chemistry* **2003,** 278, (48), 48330-48338.

177. Barril, X.; Beswick, M. C.; Collier, A.; Drysdale, M. J.; Dymock, B. W.; Fink, A.; Grant, K.; Howes, R.; Jordan, A. M.; Massey, A.; Surgenor, A.; Wayne, J.; Workman, P.; Wright, L., 4-amino derivatives of the Hsp90 inhibitor CCT018159. *Bioorganic & Medicinal Chemistry Letters* **2006,** 16, (9), 2543-2548.

178. Gopalsamy, A.; Shi, M. X.; Golas, J.; Vogan, E.; Jacob, J.; Johnson, M.; Lee, F.; Nilakantan, R.; Petersen, R.; Svenson, K.; Chopra, R.; Tam, M. S.; Wen, Y. X.; Ellingboe, J.; Arndt, K.; Boschelli, F., Discovery of benzisoxazoles as potent inhibitors of chaperone heat shock protein 90. *Journal of Medicinal Chemistry* **2008,** 51, (3), 373-375.

179. Kung, P. P.; Funk, L.; Meng, J.; Collins, M.; Zhou, J. Z. X.; Johnson, M. C.; Ekker, A.; Wang, J.; Mehta, P.; Yin, M. J.; Rodgers, C.; Davies, J. F.; Bayman, E.; Smeal, T.; Maegley, K. A.; Gehring, M. R., Dihydroxylphenyl amides as inhibitors of the Hsp90 molecular chaperone. *Bioorganic & Medicinal Chemistry Letters* **2008,** 18, (23), 6273-6278.

180. Barker, J. J.; Barker, O.; Boggio, R.; Chauhan, V.; Cheng, R. K. Y.; Corden, V.; Courtney, S. M.; Edwards, N.; Falque, V. M.; Fusar, F.; Gardiner, M.; Hamelin, E. M. N.; Hesterkamp, T.; Ichihara, O.; Jones, R. S.; Mather, O.; Mercurio, C.; Minucci, S.; Montalbetti, C.; Muller, A.; Patel, D.; Phillips, B. G.; Varasi, M.; Whittaker, M.; Winkler, D.; Yarnold, C. J., Fragment-based Identification of Hsp90 Inhibitors. *Chemmedchem* **2009,** 4, (6), 963-966.

181. Accelrys *Discovery Studio 2.5*, San Diego, California.

182. Rone, R.; Momany, F. A.; Dygert, M., Conformational studies on vancomycin using QUANTA-CHARMM. In *Smith, J. a. and J. E. Rivier*, 1992; pp 299-301.

183. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* **1998,** 102, (18), 3586-3616.

184. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **1997,** 267, (3), 727-748.

185. Meyer, E. F. J.; Karrer, A.; Radhakrishnan, R.; Trainor, D. A.; Bode, W., Differential Binding of Peptide-Analog Inhibitors to Porcine Pancreatic Elastase. *Abstracts of Papers Chemical Congress of North America* **1988,** 3, (1), BIOL 51.

186. Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L., New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *Journal of Computational Chemistry* **2003,** 24, (11), 1348-1356.

187. Fedorov, R.; Hartmann, E.; Ghosh, D. K.; Schlichting, I., Structural basis for the specificity of the nitric-oxide synthase inhibitors W1400 and N-omega-propyl-L-Arg for the inducible and neuronal isoforms. *Journal of Biological Chemistry* **2003,** 278, (46), 45818-45825.

188. Halgren, T. A., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996,** 17, (5-6), 490-519.

7. Appendix

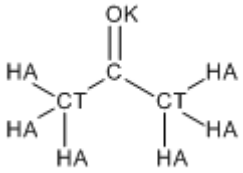
7.1. Atom Typing and Partial Charge Assignment

The preparation of ligands (solvent probes and fragments) for MCSS calculations (Chapter 4, 4.3) involved assigning atom-types from CHARMm Momany and Rone forcefield [182] and partial charges from MMFF94[188]. Here a 2-D diagram of each ligand is shown with force-field atom types as atom labels. For each ligand, atom-types and partial charges used in the calculations are also presented.

7.1.1 Solvent Probes

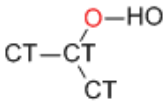
ACN

C	0.45
OK	-0.57
CT	0.06
CT	0.06
HA	0.00
HA	0.00
HA	0.00
HA	0.00
HA	0.00
HA	0.00



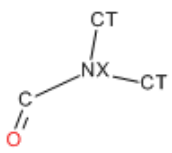
IPA

CT	0.00	HA	0.00
CT	0.28	HA	0.00
CT	0.00	HA	0.00
OT	-0.68	HA	0.00
HA	0.00	HA	0.00
HA	0.00	HO	0.40



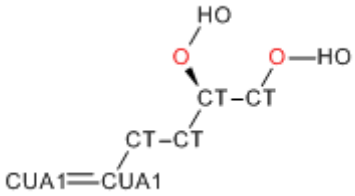
DMF

CT	0.30	HA	0.00
CT	0.30	HA	0.00
C	0.57	HA	0.00
O	-0.57	HA	0.00
NX	-0.66	HA	0.00
HA	0.00	HA	0.06



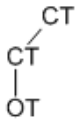
HEX

CT	0.28	HA	0.00
CT	0.28	HA	0.00
CT	0.00	HA	0.00
CT	0.14	HA	0.00
CUA1	-0.29	HA	0.00
CUA1	-0.30	HA	0.15
OT	-0.68	HA	0.15
OT	-0.68	HA	0.15
HA	0.00	HO	0.40
HA	0.00	HO	0.40



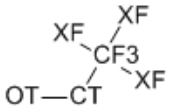
EOH

CT	0.28
CT	0.00
OT	-0.68
HA	0.00
HA	0.00
HA	0.00
HA	0.00
HA	0.00
HO	0.40



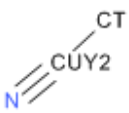
TFE

CF3	1.02
CT	0.28
OT	-0.68
XF	-0.34
XF	-0.34
XF	-0.34
HA	0.00
HA	0.00
HO	0.40



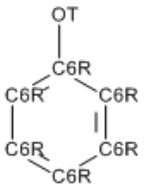
CCN

N	-0.56
CUY2	0.36
CT	0.20
HA	0.00
HA	0.00
HA	0.00



IPH

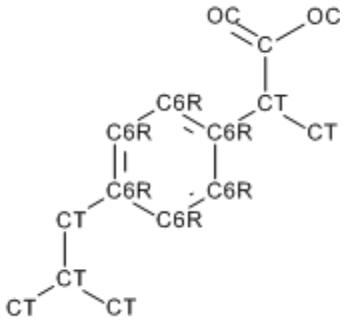
C6R	0.08	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	HO	0.45
OT	-0.53		



7.1.2 Fragment Docking Dataset

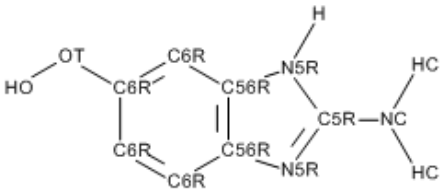
1EQG

C	0.91	HA	0.00
CT	0.14	HA	0.00
CT	0.00	HA	0.00
CT	0.00	HA	0.00
CT	0.00	HA	0.00
CT	0.04	HA	0.00
CT	0.00	HA	0.00
C6R	-0.14	HA	0.00
C6R	-0.15	HA	0.00
C6R	-0.15	HA	0.00
C6R	-0.14	HA	0.00
C6R	-0.15	HA	0.00
C6R	-0.15	HA	0.15
OC	-0.90	HA	0.15
OC	-0.90	HA	0.15
HA	0.00	HA	0.15



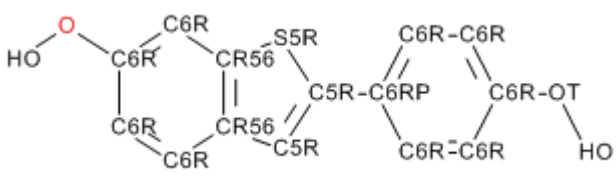
1FV9

C6R	0.08	NC	-0.88
C6R	-0.15	OT	-0.53
CR56	-0.15	HA	0.15
CR56	0.23	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	H	0.27
N5R	0.03	HC	0.40
C5R	0.27	HC	0.40
N5R	-0.57	HO	0.45



1GWQ

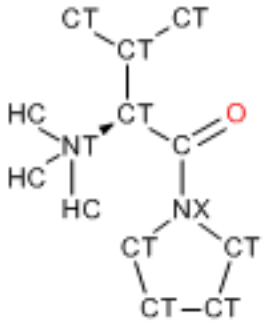
C6R	-0.15	C6R	-0.15
C6R	-0.15	CR56	0.00
C6R	0.08	C5R	-0.15
OT	-0.53	HA	0.15
C6R	-0.15	HA	0.15
CR56	0.04	HO	0.45
S5R	-0.08	HA	0.15
C5RP	-0.01	HA	0.15



C6RP	0.05	HA	0.15
C6R	-0.15	HO	0.45
C6R	-0.15	HA	0.15
C6R	0.08	HA	0.15
OT	-0.53	HA	0.15
C6R	-0.15		

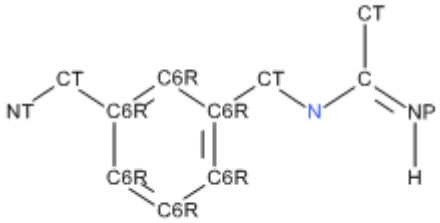
1N1M

CT	0.00	HA	0.00
CT	0.30	HA	0.00
NX	-0.66	HA	0.00
CT	0.30	HA	0.00
CT	0.00	HA	0.00
C	0.57	HA	0.00
CT	0.56	HA	0.00
CT	0.00	HA	0.00
CT	0.00	HA	0.00
O	-0.57	HC	0.45
NT	-0.85	HC	0.45
CT	0.00	HC	0.45
HA	0.00	HA	0.00
HA	0.00	HA	0.00
HA	0.00	HA	0.00
HA	0.00		



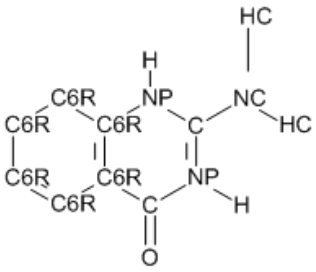
1QWC

C6R	-0.15	HA	0.15
C6R	-0.14	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	HA	0.00
C6R	-0.15	HA	0.00
C6R	-0.14	H	0.40
CT	0.51	HA	0.00
NP	-0.82	HA	0.00
C	0.44	HA	0.00
CT	0.06	H	0.40
NP	-0.85	HA	0.00
CT	0.41	HA	0.00
NT	-0.99	H	0.36
HA	0.15	H	0.36



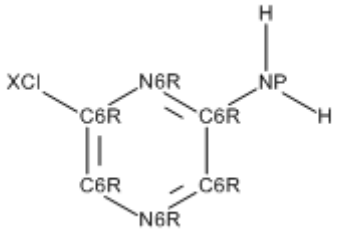
1S39

C6R	-0.15	C6R	0.09
C6R	-0.15	C6R	-0.15
C6R	-0.15	HA	0.15
C6R	0.31	HA	0.15
NP	-0.83	HA	0.15
C	1.20	HC	0.45
NC	-0.97	HC	0.45
NP	-0.86	HA	0.15
C	0.83	H	0.45
O	-0.57	H	0.45



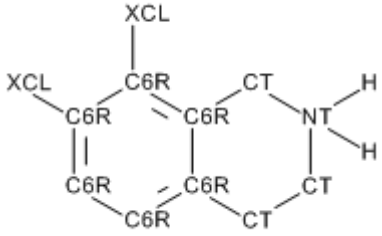
1WWC

NP	-0.90
C6R	0.41
C6R	0.16
N6R	-0.62
C6R	0.16
C6R	0.49
XCL	-0.18
N6R	-0.62
H	0.40
H	0.40
HA	0.15
HA	0.15



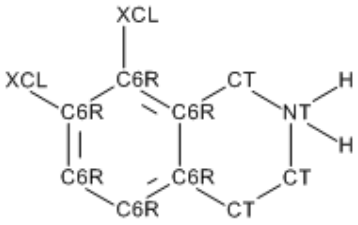
1YZ3

XCL	-0.18	XCL	-0.18
C6R	0.18	HA	0.00
C6R	-0.14	HA	0.00
CT	0.65	H	0.45
NT	-0.91	H	0.45
CT	0.50	HA	0.00
CT	0.14	HA	0.00
C6R	-0.14	HA	0.00
C6R	-0.15	HA	0.00
C6R	-0.15	HA	0.15
C6R	0.18	HA	0.15



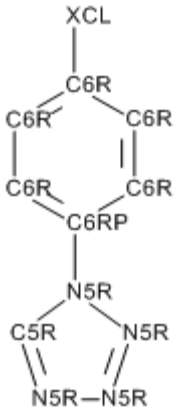
2ADU

C6RP	0.05	N5R	-0.42
C5RP	-0.20	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	HA	0.15
C5R	0.08	HA	0.15
C6R	-0.14	HA	0.15
C6R	-0.15	HA	0.00
C6R	-0.15	HA	0.00
CT	0.14	HA	0.00
N5R	0.30	H	0.27
N5R	-0.23		



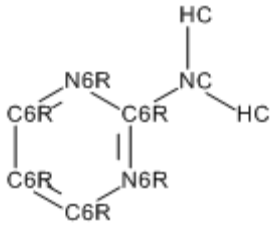
2C90

XCL	-0.18	N5R	-0.34
C6R	0.18	N5R	0.00
C6R	-0.15	N5R	-0.42
C6R	-0.15	HA	0.15
C6RP	-0.02	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	HA	0.15
N5R	0.59	HA	0.15
C5R	0.04		



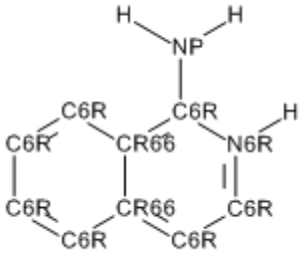
2JJC

NC	-0.90	C6R	0.16
C6R	0.72	HC	0.40
N6R	-0.62	HC	0.40
C6R	0.16	HA	0.15
N6R	-0.62	HA	0.15
C6R	-0.15	HA	0.15



2OHK

NP	-0.90	CR66	0.00
C6R	0.46	H	0.40
N6R	-0.18	H	0.40
C6R	0.21	HA	0.15
C6R	-0.15	HA	0.15
CR66	0.00	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	HA	0.15
C6R	-0.15	H	0.46



7.2. Solvent Mapping with MCSS

The following tables show results of solvent mapping with MCSS on thermolysin and elastase dataset with slightly different approach for selecting cluster representative. The cluster representative is chosen on the basis of average MCSS score of the cluster and the pose nearest to this value.

Table 7.1.Results of MCSS calculations on Elastase for solvents in generic solvent-bound structure. The predicted poses with the lowest RMSD from X-ray ($RMSD_{X\text{-}ray}$) are shown with their ranks and scores. The results after re-ranking cluster based on average scores are also indicated.

Solvent	Sub-site	Nearest cluster			Nearest cluster (re-ranked)		
		$RMSD_{X\text{-}ray}$	Rank	Score	$RMSD_{X\text{-}ray}$	Rank	Score
ACN1001	S1	2.19	31	10.87	2.19	30	10.87
IPA1001	S1	0.52	12	16.81	0.52	5	16.81
IPA1002	S4	2.12	19	14.71	2.12	14	14.71
IPA1003	S3	1.82	3	14.62	1.04	15	14.67
DMF1004	S3’	0.87	9	19.72	1.17	10	18.43
ETH1001	S1	1.98	14	15.00	1.84	20	12.06
ETH1002	S4	1.58	21	13.14	2.12	27	11.56
ETH1003	S3	1.44	11	15.21	1.43	13	13.58
ETH1004	S3’	1.96	46	10.10	1.96	46	10.10
HEX1001	S1	3.08	68	5.03	3.08	68	5.03
HEX1004	S3’	8.81	61	6.11	8.81	56	6.11
TFE1001	S1	1.63	15	11.35	1.63	8	11.35
TFE1002	S4	2.71	19	10.90	2.21	31	8.32
TFE1003	S3	0.94	2	13.17	2.21	3	12.14
TFE1008	S1’	2.38	26	9.78	2.81	23	9.42

Table 7.2. Results of MCSS calculations on Thermolysin for solvents in their native protein structures. The predicted poses with the lowest RMSD from X-ray ($RMSD_{X\text{-}ray}$) are shown with their ranks and scores. The results after re-ranking cluster based on average scores are also indicated.

Solvent	PDB	Sub-site	Nearest cluster			Nearest cluster (re-ranked)		
			$RMSD_{X\text{-}ray}$	Rank	Score	$RMSD_{X\text{-}ray}$	Rank	Score
ACN1	1FJQ	S1’	1.72	1	26.92	1.52	1	26.10
CCN1	1FJU	S1’	2.26	1	23.18	2.26	1	23.18
IPH1	1FJW	S1’	0.61	2	22.39	2.53	2	19.94
IPA1	8TLI	S1’	1.61	10	15.05	1.61	9	15.05
IPA5	8TLI	S8	2.19	17	13.41	1.63	14	13.39
IPA8	8TLI	S5	1.53	11	14.77	1.15	7	15.65
IPA9	8TLI	S2’	2.31	20	12.58	2.43	18	11.61

7.3. ΔG values calculated from MM-GB/SA method

In the following detailed results of MM-GB/SA scoring are presented for fragment-docking and HSP90 datasets. This includes the values of ΔG for the top-scoring poses in each case. All ΔG values are reported in kcal mol^{-1}

Table 7.4. Results of MM-GB/SA scoring of MCSS poses generated for fragment-docking dataset along with ΔG estimates for the X-ray pose (ΔG_{Xray}), *in situ* minimized X-ray pose ($\Delta G_{XrayMin}$) and top-scoring pose (ΔG_{Best})

PDB	$RMSD_{Xray\|XrayMin}$	ΔG_{Xray}	$\Delta G_{XrayMin}$	ΔG_{Best}	MCSS rank	$RMSD_{Xray}$	$RMSD_{XrayMin}$
1EQG	0.3	-22.52	-25.32	-25.42	1	0.3	0.1
1FV9	1.0	-9.39	-17.08	-18.45	1	2.0	1.9
1GWQ	0.5	-1.96	-23.20	-23.56	2	7.1	7.1
1N1M	0.5	-10.12	-35.39	-35.89	1	0.8	0.4
1QWC	0.8	-62.50	-95.00	-144.93	1	1.0	1.2
1S39	0.3	-34.14	-45.33	-45.63	1	0.3	0.0
1WWC	0.2	-22.77	-23.94	-23.84	22	0.2	0.0
1YZ3	0.4	-13.35	-19.73	-19.58	1	0.4	0.0
2ADU	0.5	72.61	62.59	-5.17	15	9.1	9.1
2C90	0.6	-14.90	-15.84	-15.88	2	0.6	0.0
2JJC	0.4	-8.84	-12.53	-12.55	9	0.3	0.0
2OHK	0.5	5.85	-9.34	-17.22	1	2.7	2.6

Table 7.4. Results of MM-GB/SA scoring of GOLD poses generated for fragment docking dataset along with ΔG estimates for the X-ray pose (ΔG_{Xray}), *in situ* minimized X-ray pose ($\Delta G_{XrayMin}$) and top-scoring pose (ΔG_{Best})

PDB	$RMSD_{Xray\|XrayMin}$	ΔG_{Xray}	$\Delta G_{XrayMin}$	ΔG_{Best}	MCSS rank	$RMSD_{Xray}$	$RMSD_{XrayMin}$
1EQG	0.3	-22.52	-25.32	-25.90	33	1.2	1.1
1FV9	1.0	-9.39	-17.08	-18.39	19	1.2	0.5
1GWQ	0.5	-1.96	-23.20	-34.09	4	0.6	0.0
1N1M	0.5	-10.12	-35.39	-50.29	28	0.8	0.4
1QWC	0.8	-62.50	-95.00	-98.45	4	2.5	2.6
1S39	0.3	-34.14	-45.33	-48.43	17	0.4	0.1
1WWC	0.2	-22.77	-23.94	-23.93	12	0.2	0.0
1YZ3	0.4	-13.35	-19.73	-18.13	3	0.4	0.1
2ADU	0.5	72.61	62.59	74.75	18	1.3	1.5
2C90	0.6	-14.90	-15.84	-18.14	5	5.1	5.2
2JJC	0.4	-8.84	-12.53	-12.38	6	0.3	0.4
2OHK	0.5	3.10	-10.23	-15.35	10	3.3	3.2

Table 7.5. Results of MM-GB/SA scoring of MCSS poses generated for HSP90 dataset along with ΔG estimates for the X-ray pose (ΔG_{Xray}), *in situ* minimized X-ray pose ($\Delta G_{XrayMin}$) and top-scoring pose (ΔG_{Best})

PDB	$RMSD_{Xray\|XrayMin}$	ΔG_{Xray}	$\Delta G_{XrayMin}$	ΔG_{Best}	MCSS rank	$RMSD_{Xray}$	$RMSD_{XrayMin}$
1QYE	1.0	11.11	-6.22	-10.50	2	4.1	4.1
1ZWH	2.2	116.03	11.61	-7.63	23	2.1	1.1
2CCS	0.3	51.97	-12.78	-8.66	12	5.9	5.9
2JJC	0.4	-8.84	-12.53	-12.55	9	0.3	0.0
2QF6	0.3	-12.09	-22.95	-23.01	1	0.4	0.1
2QFOa	0.9	-5.46	-10.80	-11.78	51	3.7	3.6
2QFOb	0.4	-9.60	-12.29	-16.44	11	2.3	2.1
2WI1	0.9	-5.93	-30.33	-30.11	10	0.9	0.1
2WI2	0.5	-27.13	-27.72	-33.38	1	1.0	0.9
3BM9	0.3	150.80	-28.24	-34.70	1	0.3	0.6
3EKO	0.3	267.84	-14.08	-15.17	7	5.1	4.9
3FT5	0.2	-19.18	-21.16	-23.12	2	0.8	0.7