Ligand Hydration and Bridging Waters Molecules



Akshay Sridhar

Structural Bioinformatics and Computational Biochemistry Unit

Department of Biochemistry

University of Oxford

This dissertation is submitted for the degree of

Master of Science by Research

St. Cross College

Trinity Term 2016

Abstract

Water is the most abundant molecule in biological systems and is ubiquitous in nature. Despite this prevalence, water has long been considered a backdrop for bio-processes. However, with increasing sophistication in experimental and computational methods, water has now been recognized to interact and influence the behaviour of biomolecules in complex, subtle, and essential ways. The structure of water is unique in that it is small yet able to form multiple hydrogen bonded interactions. Water molecules are thus ideally suited to bridge interactions on biomolecular surfaces. Specifically, they have been identified to play an important role in mediating protein-ligand interactions. A methodology for the inclusion of these mediating waters has often been stated to be a major challenge facing the field of rational drug design. While notable successes exist in the design of better binders using water molecules, no single method has been able completely predict the locations of water molecules and its propensity to interact with the binding ligand.

This thesis is intended to explore the possibilities of using the hydration of ligands as a pathway to study bridging water molecules in *holo* structures. To this end, a method of discretizing and scoring water-sites around ligands from their Molecular Dynamics (MD) simulation is first developed. This was then used to correlate the hydration shells of ligands with the locations of the bridging water molecules. Further, the hydration shell of ligands is used to improve WaterDock - a protocol to predict the locations of water molecules within binding sites. Finally, a comparison of MD is made with Empirical Potential Structure Refinement - a combined experimental and computational method that is often used to resolve the solvation structure of small molecules.

Acknowledgements

Right at the start, I would have to thank my supervisors Dr. Phil Biggin and Dr. Sylvia McLain for their help, guidance and patience over the course of the past two years. The meetings have always been very stimulating and have opened many avenues for the project. I am especially thankful to Phil for his help in writing up papers, compiling this thesis and most importantly guiding me in my career.

I would also like to thank Dr. Greg Ross whose previous efforts in this field have laid a great foundation for me to build and explore upon. Andrew Johnston from the McLain group has been a tremendous collaborator over the past few months never hesitating to take time off his own thesis to address my questions about EPSR and the neutron data. Being a student of SBCB at the Department of Biochemistry has also truly been the biggest widening of my scientific horizon. I will always be in awe of the quality of work presented by my lab-mates in the weekly seminars and group-meetings.

Finally, I would like to thank my parents, whose enduring encouragement and support motivates me to aspire towards greater heights of professional and personal achievement. Thank you all very much.

Table of contents

List of figures **ix**

List of tables **xv**

1 Introduction **1**

1.1 Water Molecules in *holo* proteins 1

1.2 Bridging Waters in Molecular Recognition and Drug Design . . . 5

1.3 Water Molecules in Docking 8

1.3.1 Protein-centric Waters 8

1.3.2 Ligand-centric Waters 10

1.4 Predicting the Water Sites 11

1.5 Ligand Hydration 13

2 Methods **16**

2.1 Molecular Dynamics Simulation 16

2.1.1 Initial Atomic Velocities 17

2.1.2 Force-Fields 18

2.1.3 Equations of Motion 19

2.1.4 Periodic Boundary Conditions 20

2.1.5 Long Range Electrostatic Interactions 20

2.1.6 Coupling of Temperature and Pressure 21

2.1.7 Water Models 22
2.2 Small Molecule Parameterization 23
2.2.1 General Amber Force Field 23
2.2.2 Partial Atomic Charges 24
2.2.3 Antechamber 25
2.3 Docking 25

3 Ligand Waters and their Attributes 28

3.1 Aims and Summary 28
3.2 Introduction 29
3.3 Methods 30
3.3.1 MD Simulations 30
3.3.2 Density from Trajectory and Clustering 31
3.3.3 QT Clustering Algorithm 33
3.3.4 Dataset 34
3.3.4.1 Establishing the Clustering Procedure 34
3.3.4.2 Characterizing the Water Sites 34
3.3.4.3 Ligand Hydration and Bridging Waters 36
3.4 Results 38
3.4.1 Clustering Algorithm 38
3.4.2 Characterizing the Water Sites 40
3.4.2.1 In solution 40
3.4.2.2 In Binding Complexes 42
3.4.3 Bridging Waters and Ligand Waters 44

3.4.4 Tightly Bound waters 47
3.4.4.1 TG101348 50
3.4.4.2 Indinavir 52
3.4.4.3 Cocaine 53
3.4.4.4 Sucrose 55
3.5 Conclusion and Further Work 57

4 Predicting Bridging Waters 58

4.1 Aims and Summary 58
4.2 Introduction 59
4.3 WaterDock 2.0 Pipeline 63
4.4 Methods 65
4.4.1 Identifying Functional Group Hydration 65
4.4.1.1 Methodology 65
4.4.1.2 Dataset 67
4.4.2 Identifying Energetic Cut-off 67
4.4.2.1 Methodology 67
4.4.2.2 Dataset 68
4.4.3 Validation 68
4.5 Results 70
4.5.1 Functional Group Hydration 70
4.5.2 AutoDock-Vina Energetic Cut-off 75
4.5.3 Validation 77
4.6 PyMOL Plugin 81

4.7 Conclusion and Further Work 82

5 An in-depth solvation analysis 84

5.1 Aims and Summary 84

5.2 Introduction 85

5.2.1 Empirical Potential Structure Refinement 85

5.2.2 Molecules used for comparison 86

5.2.2.1 Alprazolam 86

5.2.2.2 Cocaine and its Stereo-isomers 89

5.3 Methods 91

5.3.1 Neutron Diffraction and EPSR 91

5.3.1.1 Alprazolam 91

5.3.1.2 Cocaine 93

5.3.2 Molecular Dynamics Simulation 94

5.3.2.1 Alprazolam 94

5.3.2.2 Cocaine and its stereo-isomers 95

5.3.3 Spatial Density Analysis 98

5.4 Results 99

5.4.1 Alprazolam 99

5.4.1.1 Fitting to Experimental Diffraction Data 99

5.4.1.2 Methanol interactions with alprazolam 100

5.4.1.3 Water interactions with alprazolam 102

5.4.1.4 Atomistic Preferential Solvation 104

5.4.1.5 Comparing the bound and solution states 106

5.4.2 Cocaine 109
5.4.2.1 Fitting to Experimental Diffraction Data 109
5.4.2.2 Behaviour of cocaine from EPSR 110
5.4.2.3 Cocaine diastereoisomers in MD at infinite dilution 112
5.4.2.4 Cocaine diastereoisomers in MD at concentrated solution 116
5.4.3 Comparing MD and EPSR 119
5.5 Conclusion and Further Work 121

6 Conclusion and Future Work 123
6.1 Role of Ligand in Bridging Waters 123
6.2 Correlating MD and EPSR 124

References 126

List of figures

1.1 Cartoon representation of the the GPCR rhodopsin with the mediating water molecules 1

1.2 Cartoon representation of the BRD4(1) bromodomain with conserved water molecules 2

1.3 WaterMap results when applied to the HIV-1 Integrase 1QS4 . . . 4

1.4 Schematic illustration of varying water configuration within the OppA binding site 5

1.5 Illustration of the location of the *displaceable* water molecule in HIV-1 protease 6

1.6 Illustration of the location of the *displaceable* water molecule in Scylatone Dehydratase 6

1.7 Illustration of the intended and resultant conformations of a trisaccharide within Concanavalin A 7

1.8 Illustration of the water molecules used in docking predictions using the GOLD methodology 9

1.9 Illustration of the water molecules attached to the ligand in the MVD ligand-centric docking method 10

1.10 Illustration of the water molecule predictions using WaterDock in the AMPA bound GluR2 structure 1FTM 12

1.11 Illustration of the hydration shell of caffeine 14

1.12 Illustration of the hydration shell of proline from Neutron diffraction and MD 14

2.1 Schematic illustration of the MD simulation steps 17

3.1 The discretization workflow used to obtain sites with high water occupancies 32

3.2 Number of discrete water sites obtained at different QT clustering r_{min} values 38

3.3 Hydration shell of caffeine at r_{min} of 1.5 Å 39

3.4 Distance between the crystallographic water and the two nearest MD clusters for different values of r_{min} 40

3.5 A comparison of the number of clustered water-sites obtained from MD simulation with the total polar surface area of the ligand . . . 41

3.6 The hydration propensity of atom types calculated across the Astex and CSAR (2012) datasets 42

3.7 Scatter plot showing the propensity of MD hydration site being occupied by different protein atoms 43

3.8 Scatter plot showing the propensity of MD hydration sites being occupied by different amino acids 43

3.9 Distribution of minimum distances between a crystallographic water and a MD water-site 45

3.10 Illustration of the location of the crystallographic water and the MD ligand hydration sites 46

3.11 Scatter Plot showing displaceability of all the ligand hydration sites as a function of their occupancy scores 46

3.12 Structure of the three ligands with tightly bound water molecules 48

3.13 Scoring and illustration of the high occupancy water-sites in sucrose and cocaine 49

3.14 Overlay of bromodomain crystal structures with the inhibitor TG101348 bound 51

3.15 Overlay of the TG101348 bound water with *apo* hydration sites . 51

3.16 Interactions of indinavir in binding site of HIV-1 and HTLV-1 . . 52

3.17 Overlay of indinavir from two complexes showing the water molecule locations 52

3.18 Overlay of the indinavir bound water with *apo* hydration sites . . 53

3.19 Interactions of cocaine in binding site of GNC92H2 antibody and dopamine transporter 54

3.20 Overlay of cocaine from two complexes showing the water molecule locations 54

3.21 Overlay of the cocaine bound water with *apo* hydration sites . . . 55

3.22 Interactions of sucrose in binding site of human splicing factor and beta-lactamse 55

3.23 Overlay of sucrose from two complexes showing the water molecule locations 56

3.24 Overlay of the sucrose bound water with *apo* hydration sites . . . 56

4.1 Schematic illustration of the WaterDock pipeline 61

4.2 Schematic illustration of the WaterDock 2.0 pipeline 63

4.3 Schematic illustration of the pipeline used to identify functional groups 66

4.4 The hydration structure of five ligand polar functional groups from MD 71

4.5 Schematic illustration of the imaginary X, Y and Z axes 72

4.6 The modelled hydration structure of five ligand polar functional groups 73

4.7 The location of water-sites around different amine groups from MD 74

4.8 The modelled hydration structure of amine groups 74

4.9 MD and modelled hydration structure of hydroxyl functional group 75

4.10 Scatter plot showing the Vina docking score against the distance to the nearest crystallographic water 76

4.11 Plot showing the prediction and false-positive rates across a range of cut-off scores 77

4.12 The location of the crystallographic and predicted water sites for the six protein structures 79

4.13 Snapshot of the PyMOL plugins developed for WaterDock pipelines. 82

4.14 Snapshot of the PyMOL plugins developed for WaterDock pipelines. 82

5.1 Schematic illustration of the EPSR simulation steps 86

5.2 Molecular structure of alprazolam with the atomic labelling 87

5.3 Binding configuration of alprazolam in BRD4(1) 88

5.4 Molecular structure of cocaine with chiral carbons 89

5.5 Stereoisomers of cocaine and the atom numbering scheme 90

5.6 Binding configuration of cocaine in the GNC92H2 antibody 91

5.7 Comparison of the measured $F(Q)$ with those from EPSR and MD 99

5.8 Comparison of the measured $F(Q)$ with those from EPSR and MD in real space 100

5.9 RDF and SDM plots of methanol around the alprazolam ring structures 101

5.10 RDF and SDM plots of water around the alprazolam ring structures 103

5.11 RDF plots of methanol and oxygen around salient alprazolam atoms 105

5.12 Analysis of the mobility of the benzene ring structure of alprazolam 107

5.13 Comparison of alprazolam solvation WMA with crystallographic waters 108

5.14 Comparison of the measured $F(Q)$ with those from EPSR and MD 109

5.15 Comparison of the measured $F(Q)$ with those from EPSR and MD in real space 110

5.16 Hydration and conformation of cocaine molecules from EPSR . . 111

5.17 Analysis of the flexibility of the cocaine diastereoisomers 113

5.18 Snapshot of the most favoured location of water molecules around cocaine from infinite dilution MD 114

5.19 Hydration of the amine groups of cocaine diastereoisomers 115

5.20 SDM of the location of water molecules around the methoxy ester carbonyl O for the four stereoisomers 116

5.21 Analysis of the inter-molecular interactions and hydration of cocaine 117

5.22 SDM of the location of cocaine and pseudococaine around each other 117

5.23 Analysis of the dimerization of cocaine 118

5.24 SDM of the hydration of agglomerated and non-agglomerated cocaine 119

5.25 Analysis of the hydration effects of cocaine dimerization 120

5.26 Inter-molecular RDF plots investigating alprazolam-alprazolam interactions 120

5.27 SDM plot showing the nearest-neighbour position of diazepine ring centres around other alprazolam diazepine 121

List of tables

2.1 Weighting co-efficients of five terms included in the AutDock Vina scoring function 27

3.1 CSD small-molecule hydrated structures used to validate clustering 35

3.2 Details of the selected proteins, their apo PDB code and Resolution 37

3.3 PDB IDs of the chosen crystallographic structures from each protein family 37

3.4 RCSB code, common name and the PDB ID of ligands with 'high' occupancy hydration sites. 48

3.5 The PDB IDs of the crystal structures of ligands with high-occupancy water-sites chosen for further analysis 50

4.1 The structures used to identify consensus water-sites in the WaterDock algorithm. 62

4.2 The H-bond saturation limit enforced on the various functional groups used in the WaterDock 2.0 pipeline. 64

4.3 The dataset of structures used validate the ligand-directed Waterdock 2.0 algorithm. 69

4.4 The OppA dataset of structures used to allow comparison of new prediction protocol to AcquaAlta and the original WaterDock methodologies. 70

4.5 Individual Protein bridging water-site prediction results using the WaterDock 2.0 pipeline 78

4.6 Results of WaterDock2 with the OppA dataset 81

5.1 Cocaine bound crystallographic Structures in the Protein Database. 90

5.2 Solvent compositions of the six alprazolam solution samples. . . . 92

5.3 Solvent compositions of the six cocaine solution samples. 93

5.4 The four cocaine diastereoisomers and their PubChem CID. . . . 95

5.5 Atomic charge dispersal within the cocaine diastereoisomers. . . . 96

5.6 Cocaine stereoisomer simulation particulars. 97

5.7 Co-ordination numbers and preference ratios of alprazolam atoms from MD and EPSR 106

5.8 Comparison of the N-O_{20} bond between cocaine and allococaine. . 113

Chapter 1

Introduction

1.1 Water Molecules in *holo* proteins

In physiological conditions, proteins and other biomolecules are surrounded by water [1, 2]. With its small size and capability to act as both hydrogen bond donor and acceptor, water molecules at specific sites within proteins are often considered intrinsic to their structure and function [3–6]. For example, in the G-protein-coupled receptor (GPCR) rhodopsin, two networks of water molecules within the trans-membrane helices have been shown to mediate interactions between highly conserved residues and thus regulate activity (see Figure 1.1) [7, 8].

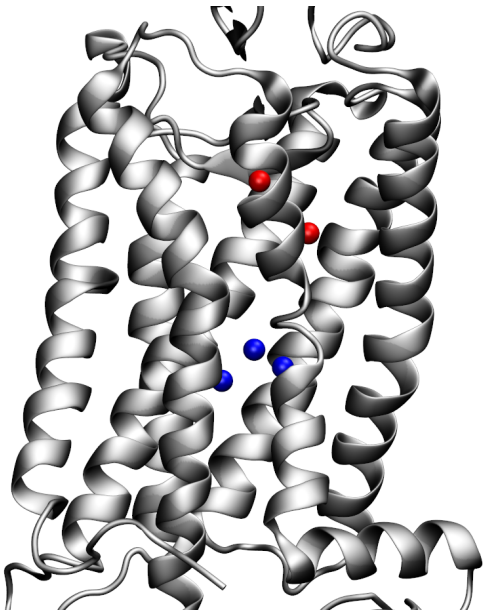


Figure 1.1 Cartoon representation of the the GPCR rhodopsin (PDB ID: **1L9H**) [8] with the sites of the mediating chains of water molecules coloured in red and blue respectively.

Although water molecules at the buried sites are exchangeable with bulk water, they can have long residence times (10^{-2} s) [9, 10] and are thus often conserved across crystallographic protein structures. For example, bromodomains are α-helical bundles that recognize acetylated side-chains in histone tails [11, 12]. Despite being substantially hydrophobic, all families of bromodomains contain an ordered chain of water molecules near the bottom of the binding site (see Figure 1.2) [13].

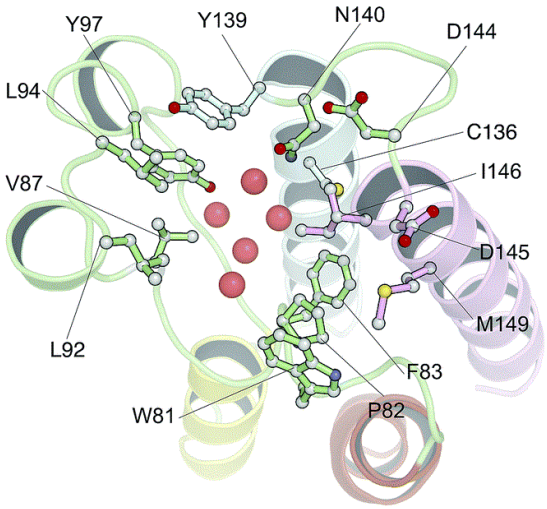


Figure 1.2 Cartoon representation of the BRD4(1) bromodomain with the key residues of the binding site labelled and the water molecules shown as red spheres. Image courtesy Aldeghi et al. [13].

The formation of a complex *holo* structure requires direct protein-ligand interactions and thus the desolvation of the protein binding site. However, a complete displacement of all water molecules rarely occurs and binding pockets have waters that interact with both protein and ligand. In fact, a comprehensive analysis by Lu et al. [14] of 392 high resolution *holo* structures found that 84% have at least one water molecule within the binding pocket. This displacement/conservation of water molecules within the binding pocket is dictated by the interplay of enthalpic and entropic contributions [15–17].

Water molecules within the protein pocket are restrained and their displacement to the bulk is thus entropically favourable. Dunitz et al. [18] calculated

this entropic gain from the removal of a water molecule to be upto 2 kcal/mol. Conversely, water molecules are capable of forming hydrogen bonds with the protein and their displacement to the bulk involves breaking these interactions and is thus enthalpically unfavourable. This fine balance of opposing factors has motivated the distinction between 'strongly' and 'weakly' bound water molecules [15, 19]. 'Strongly' bound water molecules in hydrophilic pockets interact extensively with the surrounding residues and the entropic gain from their displacement cannot overcome their energy of binding. On the other hand, the entropic gains from the displacement of 'weakly' bound waters is greater than their binding energy. This discrimination is often done either through energy calculations or a knowledge-based analysis of the water's environment.

Barillari et al. [19] used the double decoupling method together with Monte Carlo simulations to calculate the absolute binding free energies of water molecules. In their calculations of 54 water molecules from 6 proteins, 'displaceable' water molecules were on average 2.5 kcal/mol less strongly bound. A similar work by Amadasi et al. [20] scored water molecules using the Hydropathic INTeractions (HINT) force field [21] and the RANK algorithm (plausible H-bonds) [22]. A combination of the two metrics enabled a discrimination of upto 87% between displaceable and non-displaceable waters. Inhomogeneous Fluid Solvation Theory (IFST) developed by Lazaridis et al. [23, 24] uses a short Molecular Dynamics (MD) simulation to identify water distribution which in-turn is used to calculate the thermodynamics of water-binding. Beuming et al. [25] used the implementation of IFST in WaterMap [26] to identify non-displaceable water-sites and thus undruggable pockets in 27 proteins (see Figure 1.3).

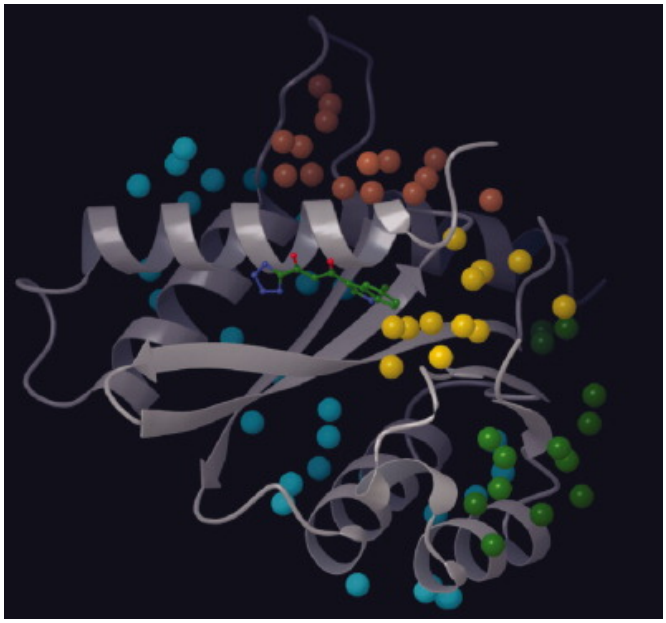


Figure 1.3 WaterMap results when applied to the HIV-1 Integrase **1QS4**. The clusters of undruggable water-sites are shown in four different colours. All four clusters are located away from the ligand binding site. Image courtesy Beuming et al. [25]

Conversely, WaterScore [27] and Consolv [28] are two knowledge based algorithms that statistically analyze the surroundings of water molecules to calculate their 'displaceability'. Consolv uses atomic density, hydrophobicity, H-bonds and crystallographic B-factor as descriptors and classifies water molecules using a k-nearest neighbour algorithm. WaterScore uses multivariate logistic regression to analyze water molecules based on their crystallographic B-factors, solvent contact area, hydrogen bond energy and number of protein-water contacts. While Consolv and WaterScore report prediction accuracies of 75% and 67% respectively, Consolv fails to predict as displaceable those water molecules that can be displaced by a polar ligand group. WaterScore on the other hand ignores all water molecules sterically displaced by the ligand. Additionally, both protocols use the Debye-Waller B-factor in their analysis which limits their applicability as not all structures accurately report B-factors of water molecules. WaterDock [29] is another statistical algorithm that combines heuristic hydrophilic, lipophilic and H-bonding scores of water molecules together with a tree based classifier to predict displaceability with an accuracy of upto 75%.

1.2 Bridging Waters in Molecular Recognition and Drug Design

'*Conserved*' water molecules within the binding pocket have long been identified to form strong interactions with both protein and ligand and in-effect act as a glue between them [30–32]. Interestingly, these bridging interactions have been demonstrated to fulfil seemingly contradictory roles in molecular recognition. A water molecule within the L-arabinose binding protein adds specificity to the receptor and allows discrimination between the binding sugars D-fucose and L-arabinose [33]. The subtle addition of a terminal methylene in D-fucose disrupts the water mediated H-bonding network and reduces binding strength. Contrastingly, water molecules in the oligopeptide binding protein (OppA) promote promiscuity and allow the binding of multiple small (3-4 residues) peptides within the same pocket [34]. Water molecules acting as a malleable medium, adopt different configurations depending on the substrate and thereby allow varied molecules to bind (see Figure 1.4).



Figure 1.4 Schematic illustration of the water configuration within OppA varying with the binding substrate. Image courtesy Ball et al. [1].

With the ubiquity and importance of water molecules in the binding of ligands, they are increasingly being considered an important facet of drug design [35–37]. A strategy often employed is the modification of a ligand to target the displacement of a water molecule to the bulk [38–40]. The entropic gain from the water's increased degrees of freedom due to displacement contributes to the free energy of binding of the drug. In the classical case of HIV-1 protease, first

generation inhibitors formed hydrogen bonded interactions with a water molecule located between the central flap structures. However, cyclical urea derivatives with a carbonyl oxygen designed to displace the water were thousand-fold stronger inhibitors (see Figure 1.5) [40]. Similarly, Chen et al. [39] designed inhibitors with a nitrile functional group to displace water from the binding site of Scylatone Dehydratase (SD), a target in the development of fungicides for rice plants. The resultant inhibitors had an increased binding potency of upto 20-fold (see Figure 1.6).

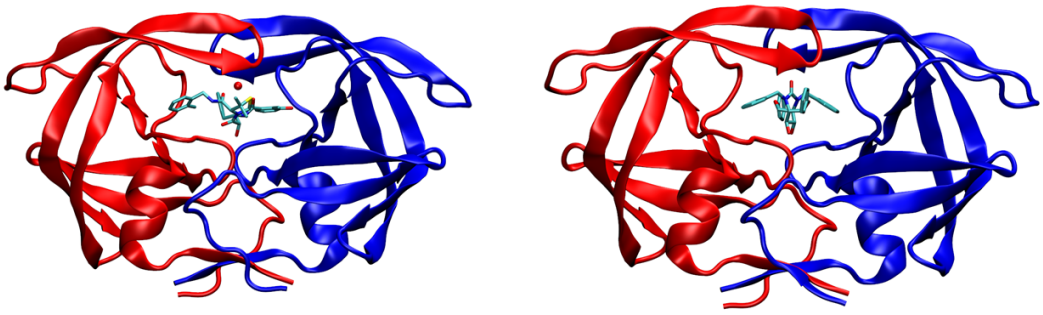


Figure 1.5 Illustration of the location of the *displaceable* water molecule in HIV-1 protease. The left pane shows the inhibitor JE2 in **1KZK** [41] that does not displace the water molecule. The right pane shows the Urea derivative displacing the water molecule in **1AJX** [42]. The two chains composing the flap structure are shown in red and blue.

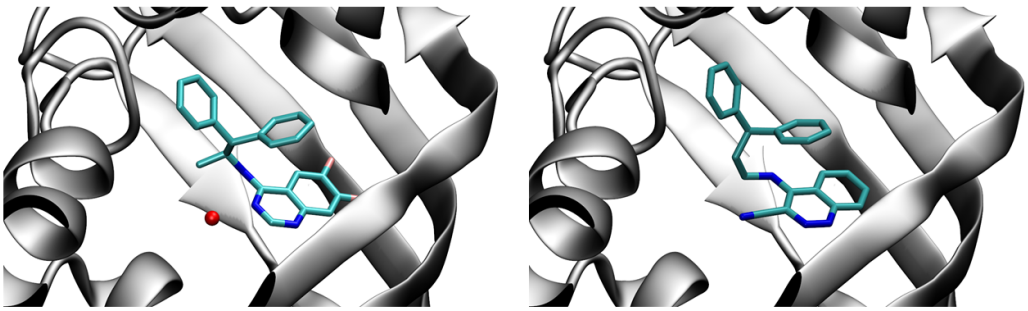


Figure 1.6 Illustration of the location of the *displaceable* water molecule in Scylatone Dehydratase. The left pane shows the inhibitor in **5STD** [43] that does not displace the water molecule. The right pane shows the nitrile derivative displacing the water molecule in **3STD** [39].

However, the rational design of ligands to displace water molecules is not always successful. Kadirvelraj et al. [44] designed a hydroxyl derivative of the

natural trisaccharide to displace a water molecule in the pocket of Concanavalin A. The synthetic ligand did not displace the water molecule but instead merely distorted its position and interactions. The analogue was thus thermodynamically less strongly bound than the natural ligand. Figure 1.7 illustrates the configuration of the natural ligand within the binding pocket in **1CVN** and the configuration of the modified ligand with respect to the water molecule in **3D4K**.

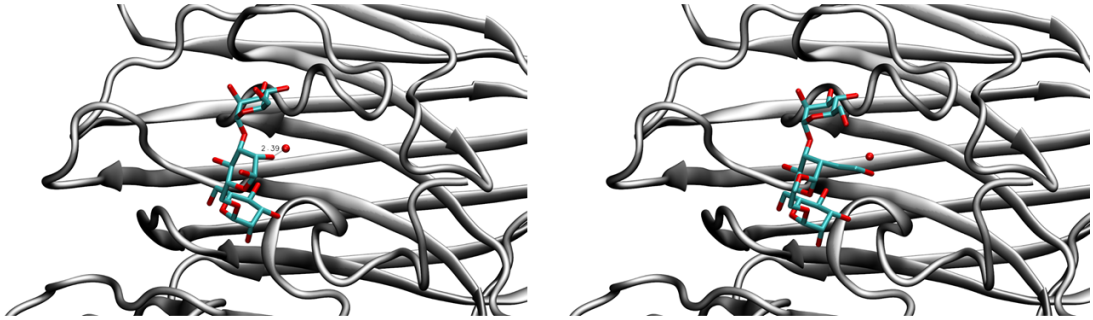


Figure 1.7 Left Pane: Illustration of the configuration of the natural ligand within the binding pocket in **1CVN** [45]. **Right Pane:** Illustration of the final configuration of the ligand analogue with respect to the water molecules in **3D4K** [44].

Water molecules have also been included in pharmacophore modelling based virtual screening methods. Pharmacophore models using protein structure do not require a prior knowledge of binding configuration [46]. The results are thus not biased by the chemical space of the previously identified binders. However, this leads the number of lead compounds identified by the screening methodology to be typically very large. Hu et al. [35] included the sites of tightly bound water molecules in pharmacophoric modelling to reduce the chemical space and have reported upto a 200 fold increase in efficiency. Similarly, Lloyd et al. [47] included water molecules when pharmacophore modelling the binding scaffolds of poly(ADP-ribose)-polymerase (PARP) inhibitors. Through the inclusion of the water molecules, they report a reduction in the chemical space of the ligand and also a more efficient screen.

1.3 Water Molecules in Docking

Molecular docking is one of the most frequently used methods in structure-based drug design due to its ability to efficiently predict the pose of ligands within the binding site [48]. Docking consists of two components: sampling of the ligand poses and the scoring of the resultant structures [49]. Despite the huge growth in the number of docking algorithms and scoring functions, no single method is able to accurately predict both the pose and affinity of probable binders [50]. With the important and multi-faceted role played by water molecules within the binding pocket, attempts have been made to improve docking performance through its incorporation. This inclusion of water molecules have been attempted through protein-centric and ligand-centric approaches.

1.3.1 Protein-centric Waters

The protein centric approach to the inclusion of water molecules in docking necessitates a prior knowledge of their location from crystallography. These water molecules within the crystallographic structures are considered non-displaceable and part of the protein. The sampling and scoring of ligand poses thus includes ligand-water and ligand-protein interactions. de Graaf et al. [51] attempted to include all X-ray water molecules in their docking studies of Cytochrome P450 and Thymidine Kinase. For Cytochrome P450, they observed improvement in prediction accuracy of 18% using AutoDock [52], 11% using GOLD [36] and 23% using FlexX [53]. In Thymidine Kinase, they observed similar improvements of 17% (AutoDock), 10% (GOLD) and 35% (FlexX).

However, this inclusion of all waters within the binding site does not allow the variation of water configuration with the ligand structure. Studies by Bellocchi et al. [54] on poly(ADP-ribose)polymerase-1 (PARP-1) in fact attributed a decrease in docking accuracy to the inclusion of a crystallographic water regardless of

ligand conformation. Verdonk et al. [36] attempted to address this issue in GOLD by allowing water molecules to be turned on/off. An 'on' water is considered '*conserved*' and included in the scoring of ligand-protein interactions. Additionally, an empirically calculated penalty representing the loss of entropic freedom is levied on the 'on' waters. An 'off' water is considered *displaced* and plays no role in ligand binding. This methodology observed an improvement of upto 20% when applied to the Astex diverse dataset (see Figure 1.8) [55]. Similar approaches have been developed for the docking programs GLIDE and FlexX.

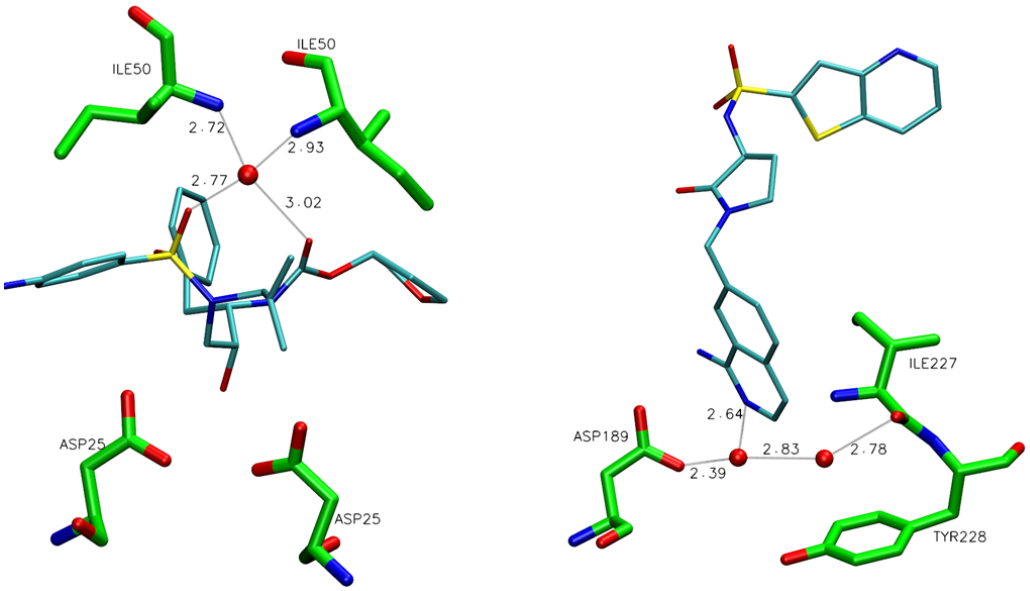


Figure 1.8 Illustration of the bridging water molecules considered in docking predictions using the GOLD methodology. Left Pane - **1HPV** [56], Right Pane - **1F0R** [57].

However, the inclusion of protein-bound waters in docking requires a prior knowledge of their location. While this is traditionally obtained from crystallography, there are associated inaccuracies. The number of water molecules identified in the structure is determined by its resolution, with the lower resolution ones having fewer [58]. Additionally, close to 10% of structures in the Protein Data Bank [59] are resolved through Nuclear Magnetic Resonance (NMR) - a method able to study only the least mobile water molecules [60]. Thus, ligand-centric approaches have been trialled as an alternate method of including water in docking.

1.3.2 Ligand-centric Waters

The ligand-centric approach to the inclusion of water molecules aims to address the issues of incomplete knowledge of protein-bound waters. Additionally, in the cases where the binding pocket itself is unknown, a ligand-centric approach allows the consideration of fewer water positions. Despite the probable advantages, limited number of methods have trialled the ligand-centric approach to docking. Lie et al. [61] incorporated the inclusion of ligand bound waters in the Molegro Virtual Docker (MVD) program [62]. Water molecules are considered attached to the H-bond capable functional groups of the ligand (see Figure 1.9) and the composite structure is docked together. Forli et al. [63] implemented a similar algorithm in the AutoDock4 force field [52] and tested the protocol on an exhaustive dataset of 1649 structures. However, performance improvement in both protocols were limited. While MVD slightly bettered the Root Mean Square Displacement (RMSD) of their predictions, it was still uable to accurately model the binding conformation. Forli et al. saw noticeable improvement only in the PARP structures within their huge dataset.

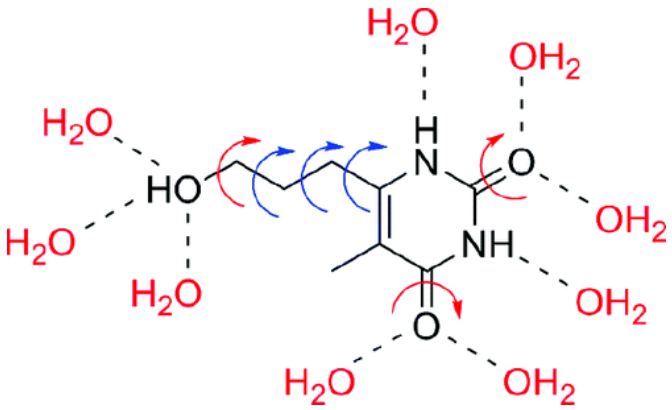


Figure 1.9 Illustration of the water molecules attached to the ligand in the MVD ligand-centric docking method. The rotatable bonds in the ligand are shown as blue arrows. Red arrows indicate additional bonds allowed to rotate to enable sampling of different water configurations. Image Courtesy Lie et al. [61].

Lemmon et al. [64] attempted a more elaborate ligand-centric approach to the inclusion of water molecules in docking using Rosetta [65]. Unlike previous

attempts where the orientation of water molecules was fixed relative to the ligand, waters were allowed to translocate and rotate moderately. This was envisaged to allow better predictions of final water location and thus better scoring. When applied to the Community Structure-Activity Resource (CSAR) dataset of 242 structures [66], predictions improved in 21, worsened in 9 and remained relatively unchanged in the rest.

1.4 Predicting the Water Sites

The inherently incomplete hydration data from crystallography makes the use of water in docking/drug design difficult. Molecular Dynamics (MD) or Monte Carlo (MC) simulations followed by a calculation of peaks in the 3D density can predict the preferential locations of water [67]. However, the method can have long convergence timescales for buried binding sites due to the time it takes for the explicit water molecules to permeate within the protein. Moreover, the use of water molecules in drug screening requires an expeditious prediction of their locations. Various 'fast-solvation' approaches have thus been trialled to swiftly estimate the locations of water within protein structures. These methods are either knowledge-based or use empirical force fields/scoring functions to predict energetically favourable water-sites.

Afanasyeva et al. [68] developed AquaBridge a method that iteratively places a water molecule on a 3D lattice, performs a local energy minimization and calculates the water's maximum number of possible H-bonds. A water molecule is considered 'structural' if it forms a minimum of two hydrogen bonds with its surroundings. When tested on 51 high resolution structures, AquaBridge could predict most crystallographic waters. However, the use of a simplistic H-bond count as a metric resulted in a high false-positive rate of upto 100% in some structures. Setny et al. [69] developed a similar iterative lattice scoring method of predicting water molecules in protein pockets. The water molecules in each

site were scored using a combination of Lennard-Jones and hydrogen bonding potentials. In addition to being slightly slower due to the more extensive scoring, the method could only predict 64% of hydration sites with a false-positive rate of 21%.

Empirical force fields initially developed for the modelling of other protein phenomenon have found use in scoring water-protein interactions. The FoldX force field [70] initially designed to score the energetic effects of point mutations was adopted by Schymkowitz et al. [71] to optimize random distributions of water molecules. The algorithm could predict upto 76% of waters in 74 crystal structures with a mean error of 0.68 Å. Similarly, Ross et al. [29] used the docking program AutoDock Vina's [72] scoring function to place and cluster predictions of water locations with an accuracy of 88% and a false-positive rate of 23%.

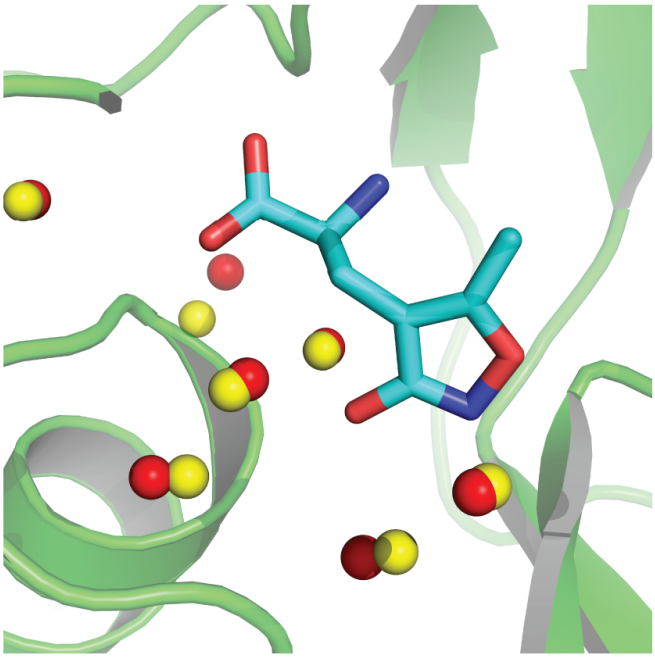


Figure 1.10 Illustration of the water molecule predictions using WaterDock in the AMPA bound GluR2 structure 1FTM. The crystallographic water molecules are show in red and the predicted sites in yellow. Image Courtesy Ross et al. [29].

AQUARIUS [73] was an early adaptation of knowledge based-methods to identify the hydration sites within proteins. A dataset of 16 high resolution crystal structures was used to 'train' the hydration structure around amino

acids. Despite the relatively small dataset, AQUARIUS could predict upto 59% of crystallographic waters within a maximum error of 1.4 Å. AcquaAlta [74] is another knowledge-based prediction algorithm that utilizes crystallographic information from the Cambridge Structural Database (CSD) [75]. The preferential orientations of water across functional groups are obtained from a search of the CSD. The intersection of these favourable geometries between the ligand and protein are then used to predict bridging water-sites.

1.5 Ligand Hydration

Computational free energy calculations are increasingly being used to judge the affinity of a drug for its targetted protein [13, 76]. Calculating the solvation free energy of a small molecule is an important first step in the binding energy predictions. This has lead work on understanding the hydration of small molecules and more importantly on the extension of Molecular Dynamics force fields to cover generic organic molecules. Lundborg et al. [77] developed STaGE, an automated methodology to parameterize and calculate the solvation energy of small molecules according to CHARMM [78], AMBER [79] and OPLS-AA [80] forcefields. Their work predicts that the parameterization of small molecules using the General Amber Force Field (GAFF) [81] and water using the TIP3P [82] model most accurately models the energetics of ligand-water interactions. While MD simulations have been used to study the solvation of a range of molecules, the hydration structures are traditionally predicted as contours (see Figure 1.11). These contours represent the most likely orientations of water molecules around the solute. However, a direct comparison of ligand hydration to *holo* bridging water-sites requires a discretization of the contours.

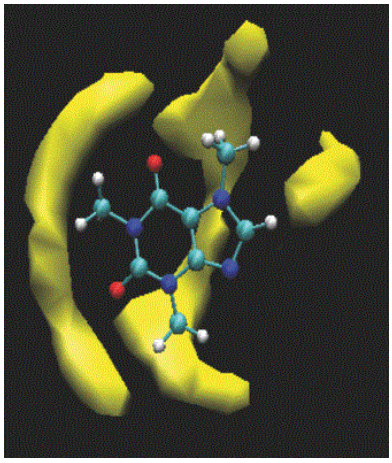


Figure 1.11 Illustration of the hydration shell contour around caffeine obtained from MD simulations. Image Courtesy Sharma et al. [83].

Neutron diffraction with its ability to discriminate hydrogen atoms is another powerful tool to study the hydration of small-molecules [84–86]. The scattering of neutrons can be used to calculate the structure factor of the mixture which in turn can be correlated to radial distribution functions (RDF). These RDFs can help understand the local solvent structure around specific atoms [87]. However, to achieve sufficient 'scattering' of solute molecules, neutron diffraction experiments must be conducted at high concentration ratios. Thus, the process has been used to study both solute-solute [88] and solute-solvent [89] interactions and the results have been generally comparable to those from MD (see Figure 1.12).

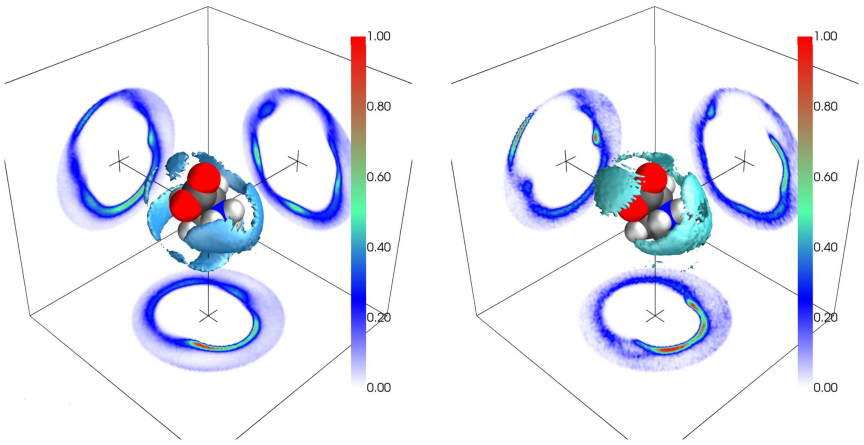


Figure 1.12 Illustration of the hydration shell of proline from Neutron diffraction (left) and MD (right). The contours show the location of the top 20% of water molecules. Image Courtesy Busch et al. [88].

Despite the breadth of work on bridging water molecules and an increasing interest in understanding the solvation of small-molecules, a correlation of the two has so far been absent. The algorithms that characterize and predict bridging water molecules are protein-directed and have largely ignored the ligand perspective. Some algorithms in fact explicitly predict limited to no role of the ligand in determining the location of bridging waters. *This work aims to model the hydration of ligands and understand its interrelationship with the important bridging waters that mediate ligand-protein interactions.*

Chapter 2

Methods

2.1 Molecular Dynamics Simulation

Molecular Dynamics (MD) Simulations serve as a powerful tool to study the behaviour of molecules *in-silico*. The simulations represent an approach to statistical mechanics with the aid of computers. In atomistic MD simulations, the system is assumed classical (ignoring quantum effects) and all atoms are represented as hard spheres [90]. Simulations begin with an initial structure of the atoms (position (r), velocity (v), connectivity) and forcefield parameters. The forces on all atoms in the system are then individually computed using the force-field. The calculated forces are then used to compute the changes in atomic displacement and velocity over a small time interval. These steps are then repeated iteratively to define the motion of atoms and thereby the behaviour of the system (see Figure 2.1).

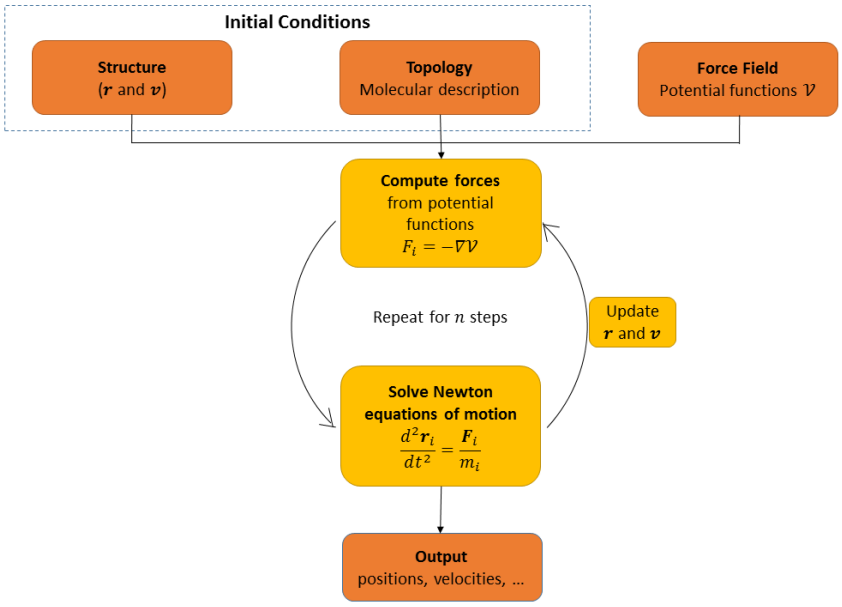


Figure 2.1 Schematic illustration of the MD simulation steps

2.1.1 Initial Atomic Velocities

The initiation of Molecular Dynamics simulations requires the position (r) and velocity (v) to be defined. The initial velocity of atoms in the system are generated randomly using the Maxwell-Boltzmann function. This function calculates the probability of an atom possessing a velocity (v) at temperature T and is described as

$$p(v_i) = \sqrt{\frac{m_i}{2\pi kT}} \exp\left(\frac{m_i {v_i}^2}{2kT}\right) \tag{2.1}$$

where v_i is the velocity of the particle, k is the Boltzmann constant, m_i is the mass of the particle, T is the temperature and p is the probability. Generating velocities from the Maxwell-Boltzmann distribution ensures that the total momentum of the system is approximately zero.

2.1.2 Force-Fields

The force-fields used in the simulations are empirical potential functions that describe the forces acting on a particle due to their interactions [91]. The total potential energy of interactions ν_{TOT} can broadly be broken up by the nature of inter-particle interactions - bonded and nonbonded.

$$\nu_{TOT} = \nu_{bond.} + \nu_{non-bond.} \tag{2.2}$$

The contributions of bonded interactions include bond stretching (ν_{bond}), bond-angle bending (ν_{angle}), improper and proper dihedral rotations (ν_{improp} and ν_{prop}). Harmonic functions are used to describe the bond-stretching, bond-bending and improper dihedral potentials. On the other hand, the proper dihedral rotations are modelled using periodic cosine functions. The total bonded interactions can thus be expressed as

$$\nu_{bond.} = \sum_{bonds} \frac{k_b}{2}(b - b_0)^2 + \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_0)^2 + \sum_{improp} \frac{k_\xi}{2}(\xi - \xi_0)^2 + \sum_{prop} k_\varphi \left(1 + \cos\left(n - n_0\right)\right) \tag{2.3}$$

where b_0 is the equilibrium bond length, θ_0 is the equilibrium bond angle, ξ_0 is the equilibrium dihedral angle, n_0 is the most favourable angle and k_b,k_θ,k_ξ,k_φ are force constants.

Non-bonded interactions ($\nu_{non-bond.}$) are described through electrostatic interactions ($\nu_{elec.}$) modelled using Coulomb's potential and van der Waals interactions (ν_{vdw}) modelled using the Lennard-Jones (LJ) potential. The total non-bonded interactions can thus be expressed as

$$\nu_{non-bond.} = \sum_{i,j}^{n} \frac{1}{4\pi\epsilon_0}\left(\frac{q_i q_j}{\epsilon_r r_{ij}}\right) + \sum_{i,j}^{n} 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} + \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{2.4}$$

where ϵ_{ij} is the strength of LJ interaction at equilibrium distance σ_{ij}, ϵ_0 is the permittivity of free space, ϵ_r is the relative permittivity of the medium and q is the partial charge on the particle.

2.1.3 Equations of Motion

The computation of the time evolution of the positions and velocities for a system of particles first requires the calculation of forces from potential-energy functions. The force (F) acting on a particle is the negative gradient of the net potential energy of the particle.

$$F = -\nabla \nu_{TOT} \tag{2.5}$$

Subsequent to the force calculations, the motion of particles are found through the integration of Newton's equation of motion.

$$F_i = m_i \frac{d^2 r}{dt^2} \tag{2.6}$$

However, with the velocity (v) of particles changing between time-steps, numerical methods like the verlet and leap-frog algorithms are used to more accurately integrate Newton's equations. The leap-frog algorithm [92] used in this work first calculates the velocity at time $t + \dfrac{\Delta t}{2}$

$$v_i\left(t + \frac{\Delta t}{2}\right) = v_i\left(t - \frac{\Delta t}{2}\right) + \frac{F_i(t)}{2m_i}\Delta t \tag{2.7}$$

where Δt is the integration time-step. This half time-step velocity is used to calculate the displacement (r) of the particle.

$$r_i\left(t + \Delta t\right) = r_i(t) + v_i\left(t + \frac{\Delta t}{2}\right)\Delta t \tag{2.8}$$

Finally, the velocity at time-step t is calculated as the average of two half time-step velocities.

$$v_i(t) = \frac{1}{2}\left(v_i\left(t - \frac{\Delta t}{2}\right) + v_i\left(t + \frac{\Delta t}{2}\right)\right) \tag{2.9}$$

2.1.4 Periodic Boundary Conditions

Molecular Dynamics simulations are normally performed in a box which encloses the system of interest in water. However, this leads to artefacts in the forces experienced by particles at the edges of the box. To overcome this, periodic boundary conditions (PBC) are applied where the system is treated as a lattice with infinite copies of itself extending in all directions. A particle in such a system that leaves a cell through one face into an adjacent box is placed back into the box from its periodic image at the opposite face. Additionally the allowing of interactions across periodic boundaries allows the effect of modelling bulk solvent with a limited number of particles. While PBC can be applied with numerous box shapes like dodecahedron and octahedron, the small ligand molecules simulated in this work allows the use of simple rectangular cuboid systems.

2.1.5 Long Range Electrostatic Interactions

Electrostatic interactions (ν_{elec}) between particles in MD simulations are modelled using Coulomb's potential (Eq. 2.4). However, for large systems with a considerable number of particles, the calculation of this summation of electrostatic potentials is computationally intensive. Algorithms like the Verlet list circumvent this by setting a cut-off radius beyond which electrostatic interactions are ignored. However, the charge imbalances within the 'lists' have been shown to cause non-physical effects in simulations.

Simulations in this work use the more advanced particle-mesh Ewald (PME) [93, 94] method of computing electrostatic interactions. In this method, the electrostatic interactions are split into short-range and long-range interactions based on a cut-off distance. The closer short-range interactions are calculated in real-space as a summation of the Coulomb's potential. The long-range interactions however are calculated as a grid-like summation in Fourier space using the fast fourier transform (FFT) algorithm. All simulations in this work use a PME cut-off distance of 10 Å and a grid spacing of 0.014 Å.

2.1.6 Coupling of Temperature and Pressure

For results comparable to those from experiments, MD Simulations are usually carried out in the NPT ensemble where the pressure (P) and temperature (T) are kept constant throughout. However, particle drift and frictional forces can cause changes in the pressure and temperature of the system respectively. Coupling techniques are thus used to maintain pressure and temperature during the course of the simulation.

The Berendsen algorithm [95] is one such thermostat which maintains the temperature of the system through coupling to an external bath. This external bath at temperature T_0 is used to correct the temperature of the system at each simulation step:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau_T} \tag{2.10}$$

where τ_T is the time-constant of temperature deviation decay. With the root mean square velocity (v_{rms}) of the particles related to the temperature by $v_{rms} = \sqrt{\frac{kT}{m}}$, the velocity of particles at the end of each simulation step is scaled by a factor λ

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)} \tag{2.11}$$

Similarly, pressure in the system is maintained through the Berendsen barostat by coupling to a reference pressure P_0 using a time-constant τ_P.

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_P} \tag{2.12}$$

At every simulation step, the particle co-ordinates and box sizes are scaled by a factor μ

$$\mu = \sqrt[3]{1 - \frac{\beta \Delta t}{\tau_P} (P_0 - P)} \tag{2.13}$$

where β is the isothermal compressibility of the system.

2.1.7 Water Models

Water models in simulation are designed with a view to balance accuracy and computational efficiency. Three point models with charges centered on each of the H-O-H atoms manage to reproduce several key factors of bulk water like density, heat of vaporization and the O-O radial distribution function. More comprehensive four and five point water models improve the prediction of additional characteristics like dielectric constant, rotational correlation times and long-range water structure albeit at a greater computational cost. Simulations in this work primarily use the three point TIP3P and SPC/E water models.

Parameterized by Berendsen et al. [96], the SPC/E model describes water as three point charges with a q_e of 0.4238 and -0.8476 on hydrogen and oxygen respectively. Additionally, the oxygen-oxygen intermolecular interactions are modelled using the Lennard-Jones potential with values of $A = 0.629\, kcal\, \text{Å}^{12}\, mol^{-1}$ and $B = 625.5\, kcal\, \text{Å}^{6}\, mol^{-1}$. The model assumes that the HOH shape is an ideal tetrahedral and has a bond angle of 109.5°. Additionally, the H-O distance in water is set as 1.0 Å.

The TIP3P water model [82] on the other hand is slightly more detailed. The point charges on the hydrogen and oxygen are similar to those in the SPC/E model with values of 0.417 and -0.834 respectively. The Lennard-Jones interaction potentials of oxygen are $A = 0.582\,kcal\,\text{Å}^{12}\,mol^{-1}$ and $B = 595.0\,kcal\,\text{Å}^{6}\,mol^{-1}$. However, unlike the SPC/E model, the HOH shape is modelled matching the experimental angle of 104.52°. Additionally, the H-O distance in water is modelled at 0.9572 Å to match the results of combined gas-phase electron diffraction and millimeter-wave spectroscopy experiments [97].

2.2 Small Molecule Parameterization

A range of force fields like AMBER [79], CHARMM [78], GROMOS [98] and OPLS-AA [80] have been developed for biomolecular MD simulations. However, the force fields have often been tuned for use with amino or nucleic acids in larger biomolecular systems [91]. Thus, general force fields [81, 99] have been developed with a view to supplement the limited parameter sets described in the more 'popular' force fields. While less accurate because of their adaptability to multiple functional groups, general forcefields have been previously developed and tested by various groups to predict hydration and solvation energy.

2.2.1 General Amber Force Field

All ligands in this work were parameterized as per the General Amber Forcefield (GAFF). Developed in 2004 by Wang et al. [81], GAFF is consistent in form and parameterization with the AMBER [100] descriptions of amino-acids. Atoms are first assigned one of the 57 atom-types based on its atomic number, hybridization (sp,sp^2,sp^3) and aromaticity (or ring structure). All the '*ideal*' equilibrium values and force constants of bond stretching, bond rotation and dihedral rotation are then looked up in a table of pre-calculated values. This table of parameters was

calculated using *ab-initio* QM/MM calculations on a dataset of 200 molecules. The van-der-Waals parameters for the Lennard-Jones potential are transferred over directly from those used for the same atoms in the AMBER99 force field [100].

2.2.2 Partial Atomic Charges

Charged interactions (ν_{elec}) in MD simulations are modelled using Coloumb's equation which necessitates the knowledge of q_e, the partial charge on each atomic sphere.

$$\nu_{elec.} = \sum_{i,j}^{n} \frac{1}{4\pi\epsilon_0} \left(\frac{q_i q_j}{\epsilon_r r_{ij}} \right) \tag{2.14}$$

While GAFF was initially parameterized using the Restrained Electrostatic Potential (RESP) [101] method of calculating charges, the procedure is computationally intensive and thus time-consuming. In this work, the q_e on the small-molecules was calculated using the AM1-BCC method of assigning charges. This method devised by Jakalian et al. [102, 103] calculates q_e as a combination of the AM1 [104] and BCC methods [105] of calculating atomic charge.

The AM1 atomic charges can be calculated very quickly and are based on the occupancies of atomic orbitals. However, they inadequately describe the dipole moments of molecules and thus perform poorly on their own in liquid-phase simulations. The Bond Charge Correction (BCC) method is thus used to '*correct*' the initially calculated AM1 charges. The BCC charge method is a semi-empirical approach with atomic charges parameterized using the standard least-square fitting procedure to match the RESP values on a training set.

A range of analytical work by Jämbeck et al. [106], Shivakumar et al. [107] and Lundborg et al. [77] have proposed the AM1-BCC charges and TIP3P water model method as the combination best suited for hydration calculations.

2.2.3 Antechamber

Antechamber [108], a program in the AMBER14 [109] suite was used to automate the above-mentioned small-molecule parameterization methodology. The program uses sequential atomic distance calculations to assign connectivity and thus atomic hybridization. Two atoms i and j are assigned to be connected if

$$D_{ij} \leqslant r_i + r_j + offset \tag{2.15}$$

where D_{ij} is the distance between the atoms and r_i/r_j are their atomic radii. The $offset$ is calculated as $c * D_{ij}$ where c is 0.15 if $D_{ij} \leq 1.5$ Å; 0.11 if $1.5 < D_{ij} \leq 1.9$ Å; 0.09 if $1.9 < Dij \leq 2.05$ Å, and 0.08 if $Dij > 2.05$Å.

These hybridization and connectivity assignments are used to lookup the parameters for the bonded, van-der-waals and coloumbic interactions.

2.3 Docking

Molecular docking is a computational approach to predict the preferred orientation of one molecule relative to a second when bound together to form a complex [48]. The process of docking thus has two major challenges - sampling the different probable conformations of the molecule and subsequently scoring each to identify the energetically most favourable conformation [49].

With each molecule having a total of six degrees of rotational and translational freedom, there is a huge number of probable binding modes. To significantly reduce the sampling space and thereby the computational load, docking algorithms reduce the number of rotational bonds in the complex. The receptor (protein) structure is generally considered to be completely rigid and only specific bonds within the ligand are assumed to be rotatable. AutoDock Vina [72], the docking program primarily used in this work utilizes a '*local search global optimizer*' [110]

to sample the probable conformations of the ligand while using an empirical scoring function to rank them.

The '*local search global optimizer*' [110] is an algorithm that uses a sequence of '*kicks*' and optimizations to find the global minima of a function. The initial ligand conformation is '*kicked*' onto a different conformation before being optimized to the local minima. The local minima is again '*kicked*' before being optimized again. This 'kicking' and optimization is repeated iteratively to find the global minima while avoiding getting trapped in a local minima. AutoDock Vina uses the Metropolis critera to accept/reject the '*kicks*'.

The Metropolis acceptance criteria and local optimization both use an empirical scoring function that combines intra- and inter-molecular interactions. Interactions between any two atoms i and j is a function d_{ij}, the surface distance defined as

$$d_{ij} = r_{ij} - R_i - R_j \tag{2.16}$$

where r_{ij} is the inter-atomic distance and R_i and R_j are the van-der-Waals radii of atoms i and j. The interaction energy is the weighted sum of five terms - $gauss_1$, $gauss_2$, $repulsion$, *hydrophobic* and *H-bonding*. The $gauss_1$ function is defined as

$$gauss_1(d) = e^{-(d/0.5)^2} \tag{2.17}$$

where d is measured in Angstroms. Similarly, the $gauss_2$ function is calculated as

$$gauss_2(d) = e^{-(d-3.0/2.0)^2} \tag{2.18}$$

The $repulsion$ term is d^2 if $d < 0$ and 0 if $d \geq 0$. The *hydrophobic* term is 1 if $d < 0.5$ and 0 if $d > 1.5$. Between 0.5 Å and 1.5 Å, the hydrophobic term is a linear interpolation between 1 and 0. The *H-bonding* term is 1 if $d < -0.7$ and 0 if $d > 0$. Similar to the hydrophobic term, the value of H-bonding term is also linearly interpolated between 1 and 0 for d_{ij} values between 0 Å and -0.7 Å. The

value of all terms are calculated upto a maximum r_{ij} of 8.0 Å. Table 2.1 lists the weighting co-efficients of the five terms.

Table 2.1 Weighting co-efficients of five terms included in the AutDock Vina scoring function

Scoring Term	Weighting coefficient
$gauss_1$	-0.0356
$gauss_1$	-0.00516
repulsion	0.840
hydrophobic	-0.0351
H-bonding	-0.587

Chapter 3

Ligand Waters and their Attributes

3.1 Aims and Summary

Bridging water molecules have long been identified to play a key role in mediating protein-ligand interactions. These molecules act as a 'glue' between the ligand and protein and are also capable of dictating binding specificity. However, the ligand perspective has largely been ignored in the study of bridging water molecules. In this chapter, I introduce a clustering methodology to identify and score sites within the ligand hydration shell and further correlate it with the positions of bridging water molecules. The results suggest that interactions with the ligand do influence the location of binding-site water molecules. A small subset of ligands in fact have tightly bound waters that are conserved across their binding to different protein pockets.

3.2 Introduction

Water is the most abundant molecule in biological systems and is ubiquitous in nature [1, 2]. As an integral part of living organisms, it fulfils multiple roles in determining the characteristics and behaviour of biomolecules [5, 8]. Specifically, water can be involved in protein-ligand interactions either through the formation of mediating hydrogen bonds or by being displaced by the ligand atom [111–114]. Mediated interactions are so prevalent that an analysis of 392 high-resolution crystal structures found that 84% have at-least one 'bridging' water molecule and that *holo* structures have on average 4-5 water molecules within their binding site [14]. On the other hand, the contribution of entropy from water's displacement to the bulk can improve binding affinity by upto 100-fold [40]. As discussed in Section 1.1, the identification of waters that are 'conserved' and thus bridge ligand binding is done based on a calculation of the strength of their interaction [19, 20, 23, 24] or a knowledge-based analysis of their environment [27–29].

However, both computational and knowledge-based calculations use the water molecule's interactions with the protein as the sole metric of 'displaceability' and ignore the contributions of the ligand. WaterDock [29] and WaterScore [27] in fact explicitly state limited to no role of the ligand in determining the location of the conserved water molecules. This counter-intuitively suggests a similar distribution of bridging water molecules irrespective of the binding configuration of the ligand. While docking algorithms that incorporate ligand-centric waters [61, 63] account for changes in ligand conformation, the final determination of conservation/displacement is still based on the strength of protein-water interactions.

This work aims to investigate the significance of the omission of ligand-water interactions in the prediction algorithms. Favourable sites within the ligand's hydration shell are first identified and their occupancy throughout MD simulation is

modelled as a simple metric of interaction strength with the ligand. A comparison of these interactions with the location of bridging water molecules then aims to identify the ligand's role in the process.

3.3 Methods

3.3.1 MD Simulations

The ligand structures obtained from the Protein Data Bank (PDB) [59] and the Cambridge Structural Database (CSD) [75] were verified to contain no regions of ambiguously resolved electron densities. If missing, hydrogen atoms were added to the ligand structures using the Reduce program [115]. *Antechamber* [108] was then used to assign atom-types and partial charges to the molecules based on the AM1-BCC [102, 103] method. The *parmchk* and *tleap* tools of AMBER14 [109] were then used to generate the topology of the ligand according to GAFF [81] in Amber format. The topologies were then converted to gromacs format using the *acpype* python script [116].

The ligands were solvated in a rectangular TIP3P box of water molecules with a minimum distance of any ligand atom to the box edge of 14 Å. Charged systems were then neutralized by the addition of Na^+/Cl^- ions as parameterized for Amber TIP3P water by Joung et al. [117]. Atomic overlaps were then corrected through Energy Minimization (EM) until the maximum force on any atom was 100 $kJ\,mol^{-1}\,nm^{-1}$. The systems were then equilibrated in the NVT and NPT ensembles for 1 ns each at a temperature of 300 K and a pressure of 1 bar using the Berendsen thermostat and barostat [95].

Production runs were carried out for 30 ns with a timestep of 2 fs and co-ordinates were saved every 0.6 ps resulting in 50000 frames. Temperature and pressure were maintained during the production runs using the V-rescale thermostat and Parinello-Rahman barostat [118] respectively. To maintain the

ligands in their '*active*' conformation, positional restraints of 100 $kJ\,mol^{-1}\,nm^{-2}$ were applied on all non-hydrogen atoms during the equilibration and production runs.

3.3.2 Density from Trajectory and Clustering

The three-dimensional spatial analysis of water co-ordinates were done through python scripts utilizing the *MDAnalysis* [119] and *scipy* toolkits. First, a 3-D grid with a spacing of 0.25 Å and extending upto 4 Å from all heavy atoms was placed over the ligand. The residence of the oxygen atom of water in each grid-point is tabulated across all frames of the trajectory. This tabulated grid is histogrammed to provide a 3-D occupancy density matrix. The three dimensional histogram is then clustered using the Quality Threshold (QT) algorithm to obtain discrete co-ordinates of high water residence. Figure 3.1 provides a basic illustration of this discretization procedure to obtain the water sites.

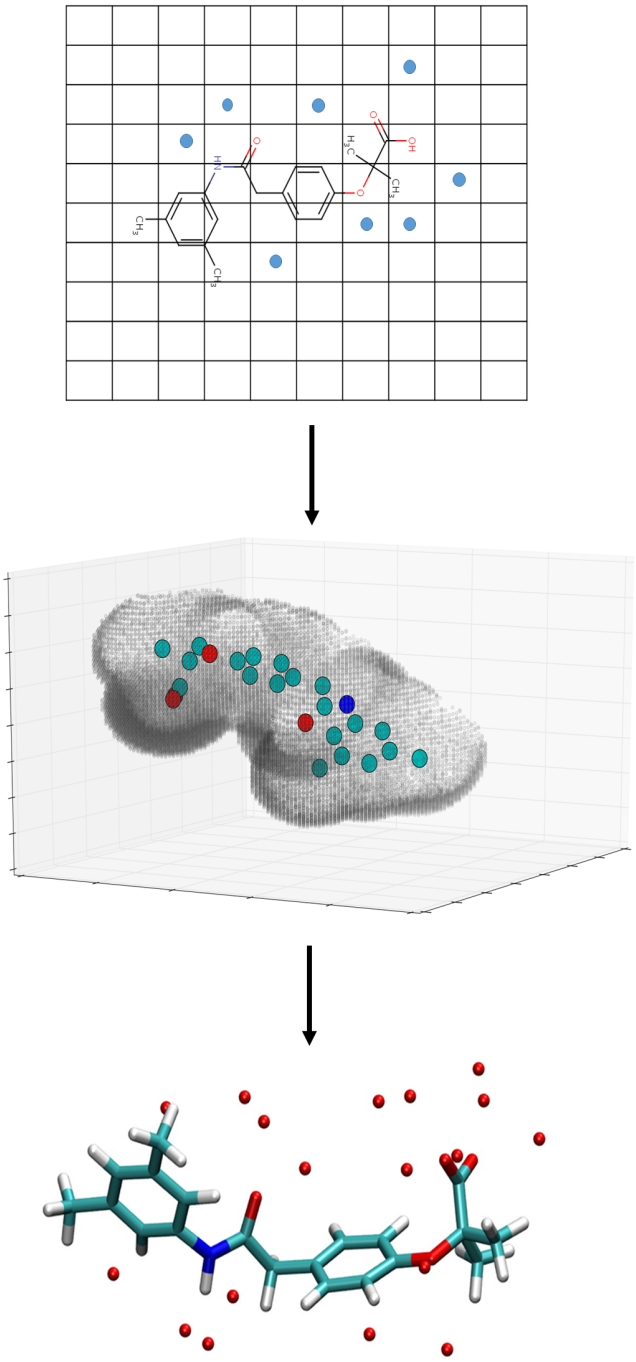


Figure 3.1 The discretization workflow used to obtain sites with high water occupancies. **Top Pane**: Ligand is placed on a grid of side 0.25 Å and the location of water is tabulated across all trajectory frames. **Middle Pane**: A 3D occupancy density matrix around the ligand is obtained through a histogram of the tabulated grid. **Bottom Pane**: The occupancy density matrix is clustered using the QT algorithm to obtain discrete water-sites.

3.3.3 QT Clustering Algorithm

First described by Heyer et al. [120] and used to group gene expression patterns, the Quality Threshold (QT) clustering algorithm has found use recently in water site analysis. *Placevent* [121], an algorithm to place important water molecules before MD simulation and *WATSite* [122], a PyMOL [123] plugin to analyze protein waters, are both based on the QT algorithm. A significant advantage of this algorithm is that a prior knowledge of the number clusters is not required. This allows its uniform application to ligands of varying size and polarity. However, a specification of r_{min}, the minimum *Euclidian* distance between two cluster sites is required.

First, a candidate cluster score ($QTscore_i$) is calculated for each grid-site as a sum of the histogrammed occupancies of its neighbouring sites.

$$QTscore_i = Occupancy_i + \sum_{|\vec{j}-\vec{i}|<\frac{r_{min}}{2}} Occupancy_j \tag{3.1}$$

The site k with the maximum $QTscore$ is selected as a discretized water coordinate. Subsequently, the occupancy of all grid-sites within $\frac{r_{min}}{2}$ of k is set to zero. The process is then repeated iteratively until the maximum $QTscore$ of a clustering iteration is less than an empirically selected cut-off.

For calculation of the cut-off, ten cubical boxes of TIP3P water with a side-length of 5 nm were simulated for 30 ns each. A random number generator was then used to select ten sites in each simulation box. The $QTscore$ of these 100 sites were then averaged and the 'cut-off' was set at 150% of this average.

3.3.4 Dataset

3.3.4.1 Establishing the Clustering Procedure

The clustering procedure and its associated distance cut-off were validated using hydrated structures from the Cambridge Structural Database (CSD) [75]. Established at the Cambridge Crystallographic Data Centre (CCDC), the CSD contains comprehensive information on small-molecules of upto 500 heavy atoms. With over 800,000 structures, the database contains crystallographic information determined through both neutron and x-ray diffraction. From the alphabetically sorted CSD, ten 'hydrated' (with waters of crystallization) drug-like organic structures were chosen while ensuring that their R-factors are low (<0.2) [124]. The residual factor or R-factor is a measure of the agreement between the crystallographic model and the experimental diffraction data [125]. It is calculated as

$$R = \frac{\sum |F_{obs} - F_{calc}|}{\sum F_{obs}} \tag{3.2}$$

where F_{obs} is the structure factor calculated from diffraction data and F_{calc} is the structure factor calculated from the model. Table 3.1 lists the chosen structures, their R-factors and the number of crystallographic waters.

3.3.4.2 Characterizing the Water Sites

With an overall aim of studying the role of ligand hydration in bridging water molecules, the distribution of the MD water-sites were characterized and compared to the distribution of water from crystallographic structures. This characterization of the hydration sites was done using a combination of the Astex and CSAR-2012 datasets.

The Astex Diverse Set [55] contains 85 high-resolution *holo* structures with pharmacologically relevant ligands. Compiled from the Protein Data Bank by

Table 3.1 CSD small-molecule hydrated structures used to validate clustering

CSD ID	Compound	R-Factor	No. of Waters
ACALDA	2,4,6-trimethyl-1,3,5-triazine	0.160	1
ACCAAH	acetylenedicarboxylic acid	0.175	1
ADODON	3,7-dihydroxy-4,14-dimethyl-7(8-9)-abeo-cholestan-8-one	0.068	2
ASPARM08	L-asparagine	0.020	1
BARBAD01	2,4,6-trioxypyrimidine	0.087	2
BOCNEM	2,4,6-trinitro benzenediol	0.042	1
CAFINE	1,3,7-trimethylpurine-2,6-dione	0.146	1
CEVSAX	2-benzyl-pentaoxa-diazacycloicosane-trione	0.056	1
CEXNIC	tris(p-(dimethylamino)phenyl)methyl chloride	0.079	3
CHEABZ	7-chloro-4-methyl-5,6-ethano-(1,4)-diazepino(1,7-a) benzimidazole	0.043	2

Astex Therapeutics, the ligands are structurally diverse and more importantly, the resolved electron density accounts for all parts of the ligand. Additionally, the bridging waters in this dataset are resolved and have been shown to be relevant to binding through docking studies [36].

The Community Structure-Activity Resource (CSAR) is a yearly compiled dataset of crystal structures aimed to provide benchmarks for the development of scoring functions and docking algorithms. The 2012 database [66] contains 242 high-resolution crystal structures and has been shown by Huang et al. [126] to possess crystallographic bridging waters capable of improving docking performance. Together, the two datasets provided 327 different ligands for hydration analysis.

3.3.4.3 Ligand Hydration and Bridging Waters

The ligand 'dependence' of bridging waters was analyzed using 54 *holo* structures of six proteins. The proteins were selected for their pharmaceutical relevance and the presence of both high resolution *holo* and *apo* crystallographic structures. The presence of many *holo* structures with a variety of ligands allows a good coverage of different functional groups and the high-resolution *apo* structure provides the co-ordinates of 'protein-bound' waters for comparison. Table 3.2 lists the proteins along with the PDB ID and resolutions of their *apo* structures.

With the inherent inaccuracies associated with crystallographic water molecules [58], the *apo*-hydration of each protein family was confirmed through MD simulations. For this, each crystal structure in Table 3.2 was stripped of its crystallographic waters and co-factors before parameterization according to the Amberff99SB-ILDN forcefield [79]. They were then simulated for 50 ns and water-sites were identified through the procedure used for the ligands and detailed in Section 3.3.3.

Table 3.2 Details of the selected proteins, their apo PDB code and Resolution

Protein	*apo* PDB ID	Resolution
BRD4	2OSS	1.35
Trypsin	1S0Q	1.02
Penicillopepsin	3APP	1.8
PIM1 Kinase	1YWV	2.0
HIV-1 Protease	4YOB	1.5
HSP-90	1AH6	1.8

The RCSB sequence similarity tool was used to obtain *holo* structures of the proteins in Table 3.2. The *holo* structures obtained were arranged by resolution and those with ambiguously resolved ligand electron densities were discarded. Table 3.3 lists the PDB IDs of the chosen crystal structures from each protein family. If more than one ligand was crystallized due to multiple protein chains, the configuration with the greater number of bridging waters was chosen for analysis.

Table 3.3 PDB IDs of the chosen crystallographic structures from each protein family

BRD4	Trypsin	Penicillopepsin	PIM1	HIV-1	HSP-90
3MXF	2AH4	2WEA	1XWS	1FGC	2IWX
3SVG	3NKK	2WEC	1R02	4ZLS	2BRC
3U5J	1UTN	2WEB	3R04	3TOH	2VW5
3U5L	3NK8	1PPL	3VBQ	1SDT	2XX2
4HBV	1UTO	1PPM	4ALW	2A1E	1ZW9
4J0R	4AB9	1PPK	4BZN	2HB3	1AM1
4MR3	4MTB	2WED	4DTK	3TOG	2FXS
4MR4	4ABB	-	4K0Y	3DJK	4CE1
4OGI	4ABA	-	4XHK	-	2BRE
4OGJ	4ABH	-	-	-	2XX5

3.4 Results

3.4.1 Clustering Algorithm

The Quality Threshold (QT) clustering algorithm employed in this work requires a specification of r_{min}, the minimum *Euclidian* distance between neighbouring cluster sites. Figure 3.2 plots the number of clustered sites for each of the small-molecules from the CSD at three different r_{min} of 1.5 Å, 2.0 Å and 2.5 Å.

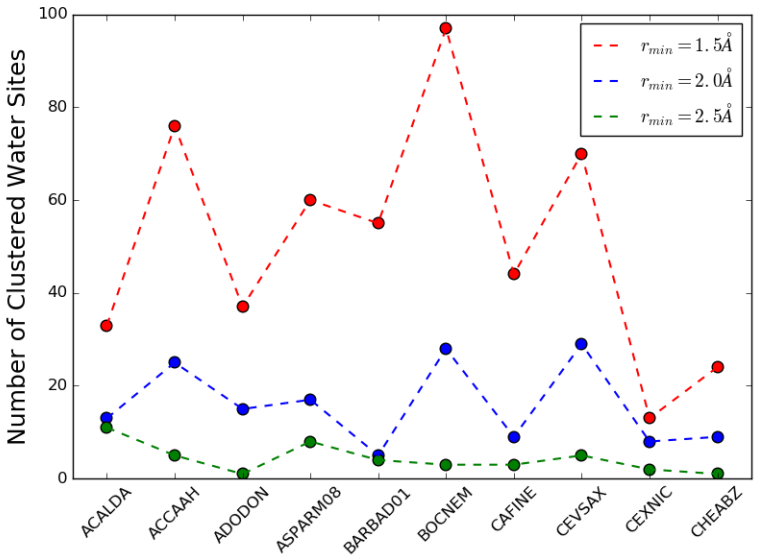


Figure 3.2 Number of discrete water sites obtained at different QT clustering r_{min} values for the CSD dataset

At small values of r_{min}, the distance between clusters is reduced and thus the number of discrete water-sites increase. With a large number of sites, the chances of missing the crystallographic water locations are reduced. However, with multiple MD clusters, correlating crystallographic water location to a specific cluster becomes challenging. Figure 3.3 elaborates this through a snapshot of the hydration shell of caffeine (Tri-Methyl-purine-dione) predicited at a r_{min} of 1.5 Å. The location of the crystallographic water (blue) is mid-way between two predicted hydration sites (red).

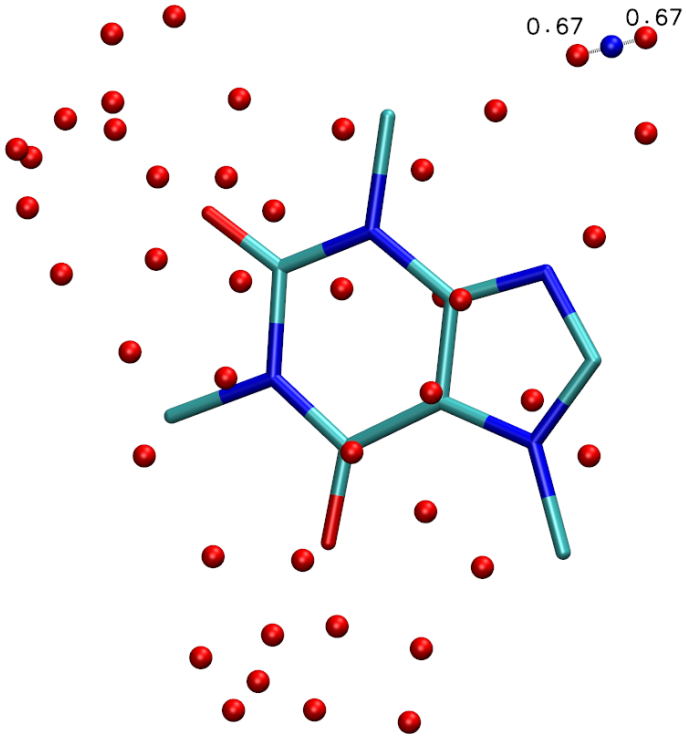


Figure 3.3 Hydration shell of caffeine at r_{min} of 1.5 Å. Crystallographic water and clustered sites are shown in blue and red respectively.

Conversely, at large r_{min}, the number of clusters are reduced and all crystallopgraphic waters are not mappable to a MD hydration site. Figure 3.4 elaborates this by plotting the *Euclidian* distance between the co-ordinates of crystallographic water and the two nearest MD water clusters at three different values of r_{min}. The distance to the nearest cluster is plotted as a red dot and the distance to the second-nearest cluster is plotted as a black 'x'. With an initial histogram width of 0.25 Å, candidate values of r_{min} are separated by 0.50 Å.

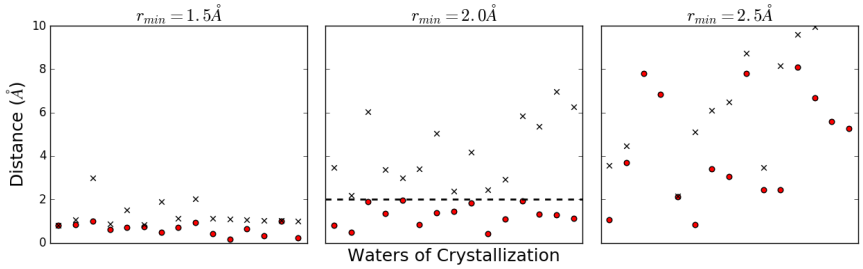


Figure 3.4 Distance between the crystallographic water and the two nearest MD clusters for different values of r_{min}. The nearest distance is plotted as a red 'o' and the second nearest is plotted as a 'x'

At 1.5 Å, the complication elaborated by Figure 3.3 is apparent with distances to the '*nearest*' and '*second-nearest*' clusters indistinguishable for many crystallographic waters. At 2.5 Å however, only 3 water molecules are mappable to a single MD hydration site. A r_{min} value of 2 Å is a suitable middle-ground with an uniform distance cut-off (*horizontal-line*) enabling all crystallographic water molecules to be mapped to a single MD cluster.

3.4.2 Characterizing the Water Sites

3.4.2.1 In solution

Figure 3.5 plots the number of water clusters predicted through MD simulation as a function of the polar surface area of the ligand. The polar surface areas were calculated using RDKit's [127] implementation of the fragment based approach devised by Ertl et al. [128]. With a R^2 value of 0.88, the number of water-sites correlates strongly with the polar surface area of the ligand. This strong correlation provides initial encouragement for the ligand-based approach. An analysis of bridging waters in high-resolution crystal structures by Lu et al. [14] found a similar correlation with a R^2 of 0.63 between the polar surface area of the ligand and the number of bridging water molecules.

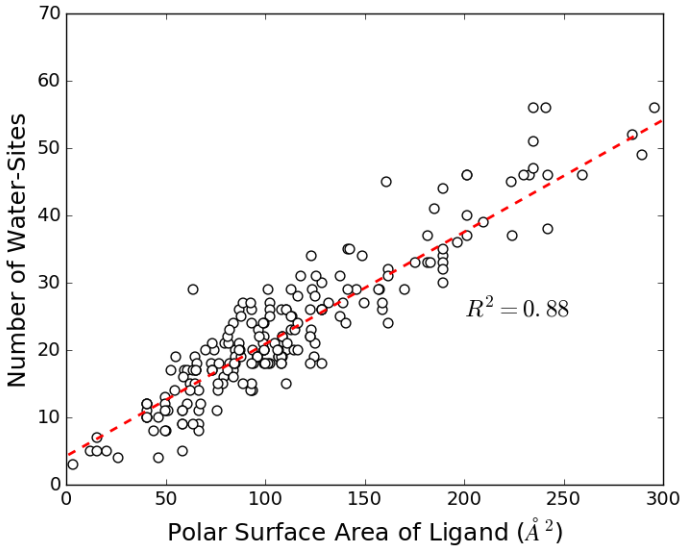


Figure 3.5 A comparison of the number of clustered water-sites obtained from MD simulation with the total polar surface area of the ligand. The R^2 value is 0.88.

Figure 3.6 further analyses the location of the cluster sites by calculating the hydration propensity of ligand atoms. First, each clustered water-site is assigned to the closest non-hydrogen ligand atom which is then treated as the primary hydration atom. The hydration propensity for each atom type is then calculated as

$$Propensity_x = \frac{site_x}{atom_x} \div \frac{site_{tot}}{atom_{tot}} \tag{3.3}$$

where $site_x$ and $site_{tot}$ are the number of cluster sites primarily hydrated by atom type x and the total number of clusters respectively. $atom_x$ is the number of atoms of type x and $atom_{tot}$ is the total number of atoms in the dataset. Oxygen atoms are most hydrated in solution followed by nitrogen, halogens and sulphur/phosphorus. Expectedly, carbon atoms are least hydrated with a hydration propensity < 0.5. A calculation by Lu et al. [14] on crystallographic waters found identical hydration propensities of oxygen, nitrogen and halogens at ∼2.9, 1.9 and 1.3 respectively. Together, the results in Figures 3.5 and 3.6, suggest a similar distribution of water molecules around ligands in solution and when bound in a complex.

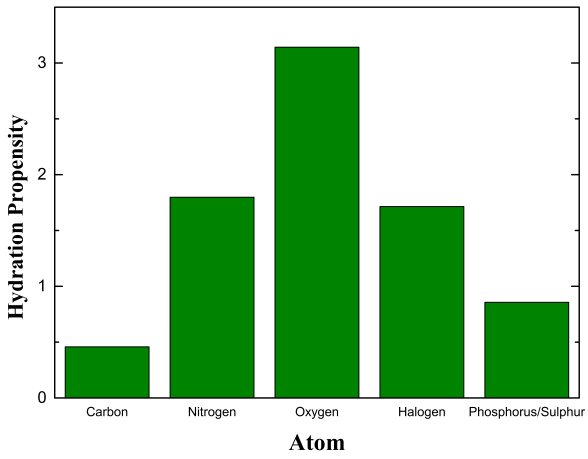


Figure 3.6 The Hydration Propensity of atom types calculated across the Astex and CSAR (2012) datasets.

3.4.2.2 In Binding Complexes

Figures 3.7 and 3.8 investigate the role of water-ligand interactions in determining the ligand's binding conformation. The ligands along with their hydration shells were mapped onto the crystallographic structure and the closest protein atom within 1.5 Å of a water-site was considered the '*displacing*' atom/residue. The clustered $QTscore$ used to cluster the water-sites measures their occupancy throughout the MD trajectory and was thus considered as a metric of ligand-water interactions.

In Figure 3.7, as the 'occupancy' ($QTscore$) of water-sites increase, the probability of being displaced by a protein oxygen atom increase and vice-versa for carbon. However, the probability of a protein nitrogen atom occupying the water-site remains constant across occupancy scores. During the development of WaterDock, Ross et al. [29] conducted a similar but inverse analysis on the displacement of protein-bound waters by ligand atoms. A similar trend was observed with oxygen most-likely and carbon least-likely to displace protein waters. Interestingly, the propensity of displacement by nitrogen atoms remained constant like in the current analysis.

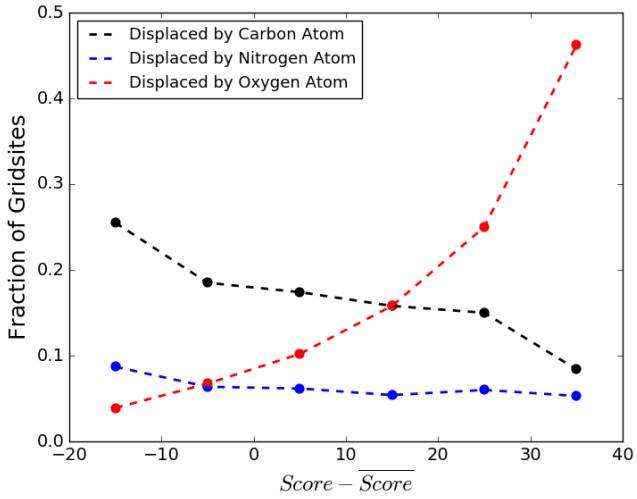


Figure 3.7 Scatter plot showing the propensity of MD hydration site being occupied by protein C(black), N(blue) and O(red) atoms as a function of the occupancy '*Score*'.

In Figure 3.8, MD water-sites across all 'occupancies' are preferentially displaced by the charged amino-acids arginine, histidine, lysine, aspartic acid and glutamic acid. However, as the 'occupancy' of the hydration site increases, its propensity to be displaced by amino acids with polar side-chains (serine, threonine, asparagine, glutamine) also increase. Conversely, the propensity decreases with occupancy for the eight amino acids with non-polar side-chains (alanine, valine, isoleucine, leucine, methionine, phenylalanine, tyrosine, tryptophan).

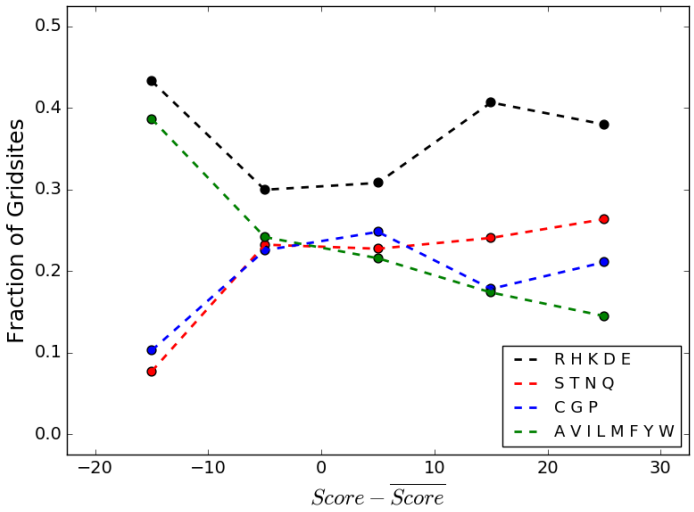


Figure 3.8 Scatter plot showing the propensity of MD hydration sites being occupied by different amino acids as a function of the occupancy '*Score*'.

The two Figures 3.7 and 3.8 together suggest that a ligand hydration site's occupancy can plausibly be an indicator of the binding configuration. However, despite the notable trends with respect to the displacing atom/residue, the total 'displaceability' remains constant across site occupancies. This total fraction of gridsites that are occupied by protein atom/residue remains invariant at $\sim$0.6 regardless of the strength of water-ligand interactions. While a direct comparison to bridging waters cannot be made due to the insufficently resolved waters in the dataset, the initial results support the traditional hypothesis that ligand-water interactions play a minimal role in determining conservation/displacement.

3.4.3 Bridging Waters and Ligand Waters

The structures in Table 3.3 were compiled to provide a dataset of high-resolution crystal structures whose bridging water sites can be identified with reasonable accuracy. 'Bridging' waters were defined as those water molecules that are within 3.4 Å of both a non-hydrogen ligand and protein atom. The 54 *holo*-structures together had 146 crystallized bridging water molecules. The MD hydration shells of the 54 ligands were mapped onto the binding configuration and the distance to the nearest MD water-site was calculated for each crystallographic water. Figure 3.9 plots this distribution of minimum distances upto 2Å - the initially identified clustering cut-off (r_{min}).

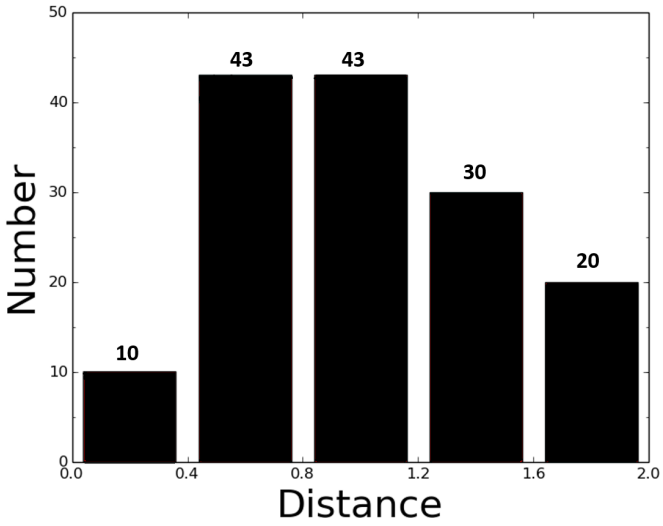


Figure 3.9 Distribution of minimum distances between a crystallographic water and a ligand hydration site mapped from MD simulation.

Within the established QT clustering r_{min}, the locations of all 146 crystallographic waters are mappable to a ligand hydration site. Figure 3.10 illustrates this by displaying a snapshot of the predicted hydration structure of the BRD4(1) inhibitor alprazolam in **3U5J** [129]. The crystallographic waters are shown in red and the MD water sites in blue. This initial result that the location of bridging waters are a subset of ligand hydration sites suggests that configuration of the ligand molecule cannot be completely discounted. For water molecules to be conserved within the binding pocket, favourable interactions with ligand atoms seems necessary.

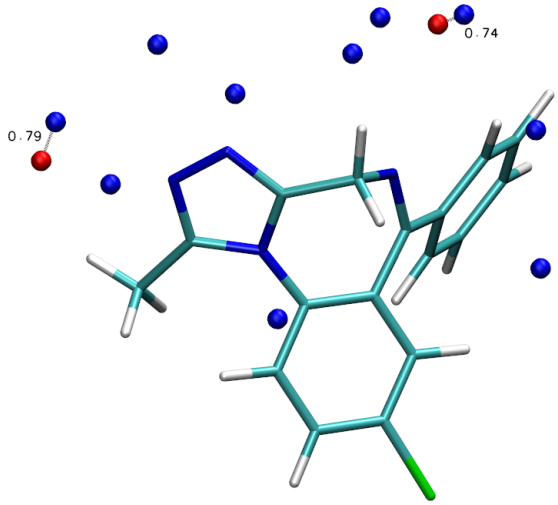


Figure 3.10 The location of the crystallographic waters and hydration MD sites of BRD4(1) inhibitor alprazolam in **3U5J** [129]. The crystallographic waters are shown in red and the MD water sites in blue.

Figure 3.11 further investigates the role of ligand-water interactions in the water-site's '*conservability*'. For each of the 1950 ligand hydration sites obtained from MD simulations of the structures in Table 3.3 and the Astex dataset, the distance to the nearest crystallographic water was calculated. The water-site was considered 'conserved' if this distance was less than 2.0Å and 'displaced' otherwise.

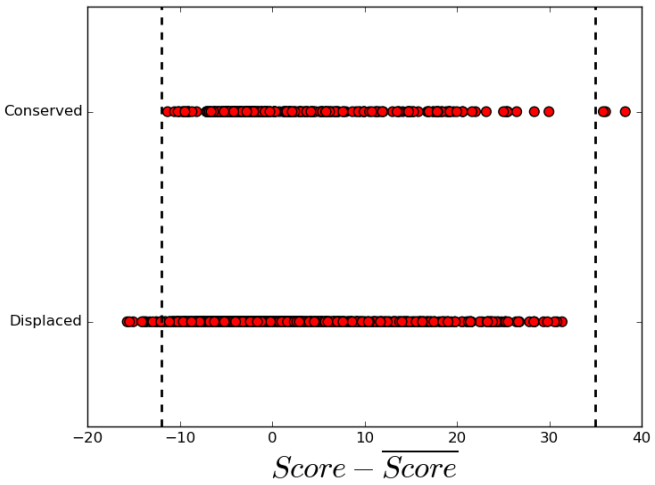


Figure 3.11 Scatter Plot showing displaceability of all the ligand hydration sites from structures in Table 3.3 and the Astex Dataset as a function of their occupancy scores.

A visual inspection of the scatter plot suggests three plausible conculsions with regards to the role of ligand in determining bridging water locations.

1. Favourable interactions with ligand atoms is a requisite for the conservation of the water-site. A similar conclusion might be drawn from the results in Figure 3.9 where all crystallographic waters were mappable to MD sites with an occupancy of at least 150% of bulk. However, Figure 3.11 suggests that this cut-off might in fact be higher with none of the 37 lowest occupancy (< -12) water-sites conserved. Hence, merely protein-water interactions cannot be stated to completely dictate the location of bridging water molecules in the binding pocket.

2. Protein interactions dictate conservation/displacement only for sites with moderately favourable interactions with the ligand. This is elaborated in Figure 3.11 where sites with occupancy scores between the two parallel lines are equally displaced and conserved. These results also correlate with those in Figures 3.7 and 3.8 where the total fraction of sites that were displaced remained approximately constant at $\sim$0.6 across a range of occupancy scores. Hence, the hypothesis of prediction algorithms like WaterScore/WaterDock that protein interactions play the deterministic role is only valid while the sites have favourable interactions with the ligand.

3. Water-sites with exceptionally favourable interactions with the ligand and thus high occupancy scores (> 35 in Figure 3.11) are always conserved. While Figure 3.11 only had three such sites, their occupancy scores from MD simulation were comparatively much higher than the others.

3.4.4 Tightly Bound waters

Table 3.4 lists the RCSB code, common name and the PDB ID of the three ligands plausibly identified with tightly bound water molecules in their active conformation. Figure 3.13 shows snapshots of the three ligand molecules along with the location of the conserved high occupancy hydration site. In all three

cases, the water-sites are subtended by hydrogen bonds with multiple ligand N/O atoms.

Table 3.4 RCSB code, common name and the PDB ID of ligands with 'high' occupancy hydration sites.

RCSB code	Common Name	PDB ID
FSN	-	1OYT
2TA	TG101348	4OGJ
MK1	Indinavir	1SDT

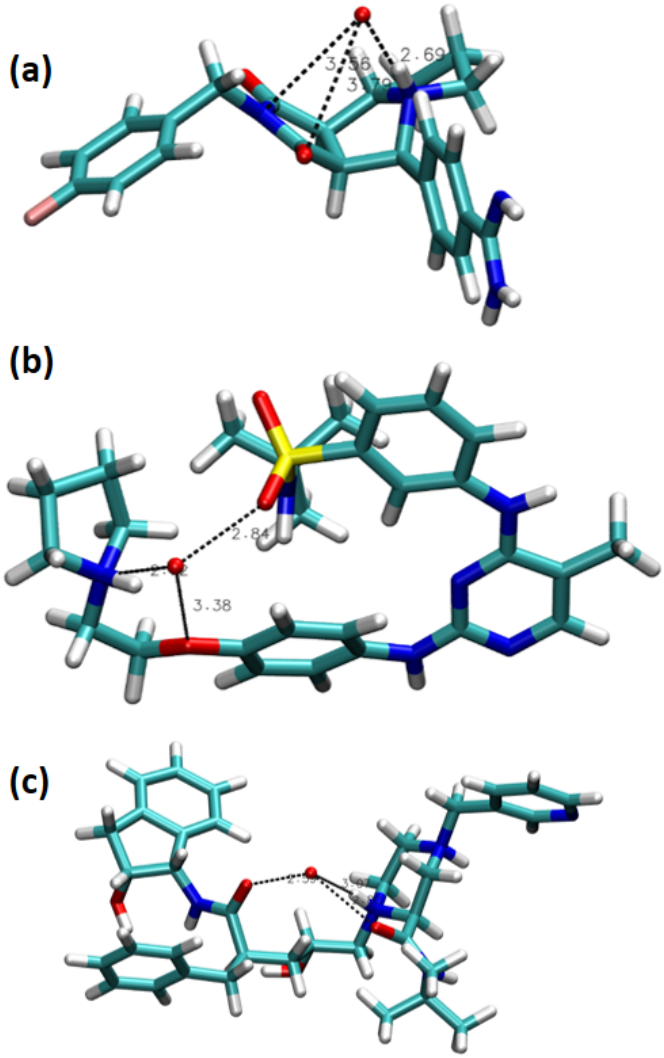


Figure 3.12 Structure of the three identified ligands with conserved high-occupancy (> 35) water sites. The location of the water sites are shown as red spheres. **(a)** FSN - 1OYT. **(b)** 2TA - 4OGJ. **(c)** MK1 - 1SDT.

To identify more such tightly-bound hydration sites and also validate the scoring metric, the previously simulated CSAR dataset was searched for a hydration site with a similarly high occupancy score (> 35). Of the 4378 predicted hydration sites from the 242 ligands, only 1 molecule (sucrose, **2ADD** [130]) was as highly scored. Despite the relatively low resolution of the structure **2ADD** (2.5 Å), a water molecule crystallized only 0.8 Å from this high 'occupancy' site.

Additionally, work with experimental collaborators[1] suggested cocaine as another molecule with a highly localized hydration density. It was thus simulated and scored according to the same methodology and the experimentally predicted water-site had an occupancy score of 35.6 in MD. Figure 3.13 shows the scoring plot of Figure 3.11 while including the sucrose and cocaine high occupancy sites. It also shows snapshots of the locations of the tightly-bound sites in sucrose and cocaine.

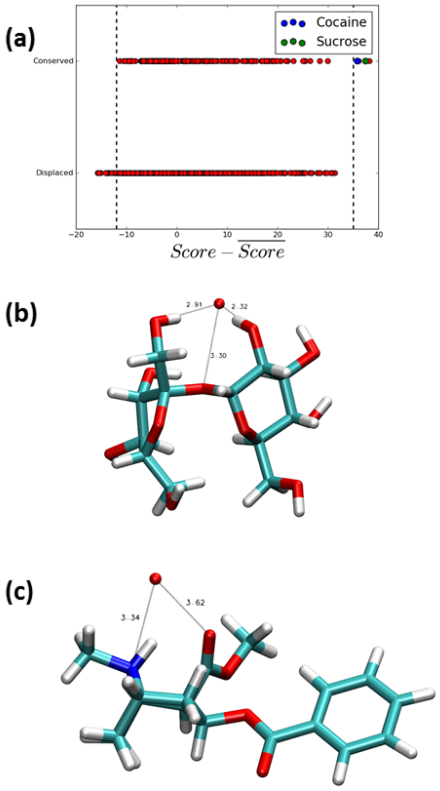


Figure 3.13 (a) Scoring metric plot of Figure 3.11 while including the sucrose (green) and cocaine (blue) high-occupancy sites. **(b)** The location of the high occupancy site in Sucrose and its associated H-bonds. **(c)** The location of the high occupancy site in Cocaine and its associated H-bonds.

[1] See Chapter 5

To further corroborate the relationship between high occupancy water sites and bridging waters in complexes, the Protein Data Bank was searched for high resolution *holo*-structures containing the five ligands. Table 3.5 lists two crystal structures of each ligand chosen for further analysis. The ligand FSN was a novel ligand and did not have a second crystallographic *holo*-structure.

Table 3.5 The PDB IDs of the crystal structures of ligands with high-occupancy water-sites chosen for further analysis

2TA	MK1	COC	SUC
4OGJ	3WSJ	4XP4	1YLJ
4PS5	1C6Y	1I7Z	3GIX

3.4.4.1 TG101348

The ligand 2TA is a selective Janus Kinase 2 (JAK2) inhibitor and is clinically termed **TG101348** [131]. It is currently under clinical trials for the treatment of Myelofibrosis and hematopoietic neoplasm [132, 133]. However, it has also been shown to interact and bind strongly with the bromodomain and extraterminal (BET) family of proteins. Filippakopoulos et al. [129] and Ember et al. [134] separately crystallized structures of the first bromodomain of human BRD4 in complex with the inhibitor. Figure 3.14 shows an overlay of the two *holo* structures and the water molecule conserved across both structures.

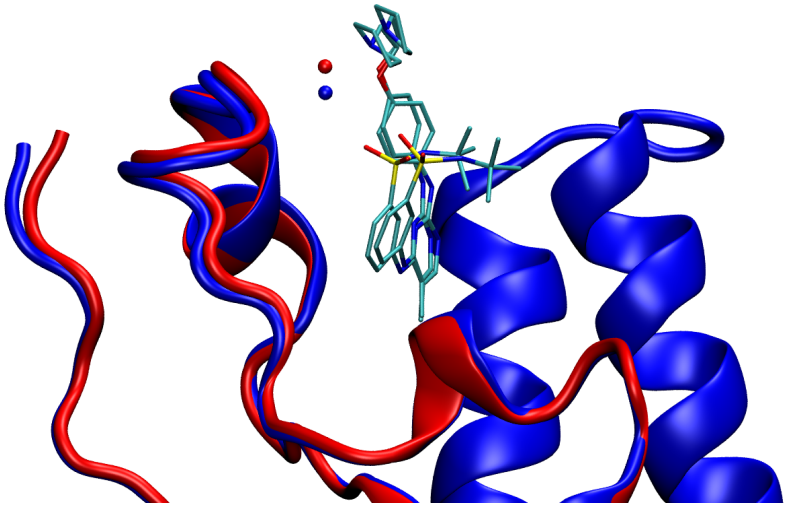


Figure 3.14 Overlay of the two bromodomain crystal structures **4OGJ** [129] and **4PS5** [134] with the JAK2 Inhibitor TG101348 bound. The conserved water molecules across the two complexes are shown as red and blue spheres.

To further confirm the ligand dependence of this water, the *apo*-structure of the BRD4(1) was simulated and high occupancy water-sites within the binding pocket were identified through clustering. Figure 3.15 overlays the *apo*-protein hydration sites with the ligand bound water molecule. The ligand-bound water molecule is located a minimum of ∼4Å from any *apo* hydration site.

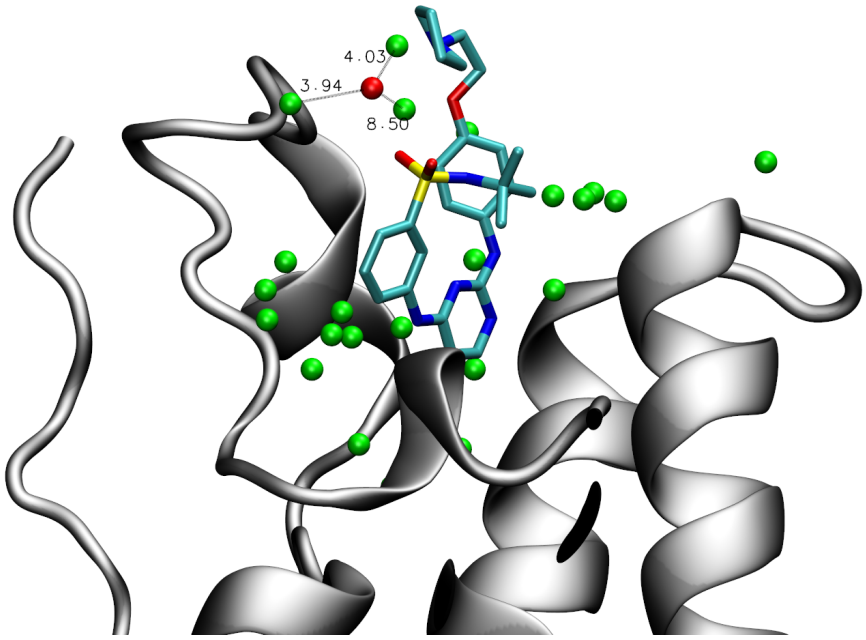


Figure 3.15 Overlay of the water bound to the TG101348 inhbitor (**4OGJ**) with the locations of *apo* hydration sites from MD. The hydration sites of the *apo* protein from simulation are shown in green and the location of the tightly-bound crystallographic water from the *holo* structure 4OGJ is shown in red.

3.4.4.2 Indinavir

Ligand MK1 is the anti-retroviral drug Indinavir (Crixivan®) and is used in the treatment of HIV/AIDS [135]. However, it has an extremely short residence-time in the body and has been shown to promote the formation of drug-resistant mutations [136]. In addition to HIV, Kuhnert et al. [137] crystallized Indinavir in complex with the Human T-lymphotropic Virus - 1 Protease (HTLV-1). Though from the same family, the binding site of Indinavir have slightly different interactions as shown in Figure 3.16. Despite this, the two structures both have water molecules crystallized at the high-occupancy site predicted by MD (see Figure 3.17).

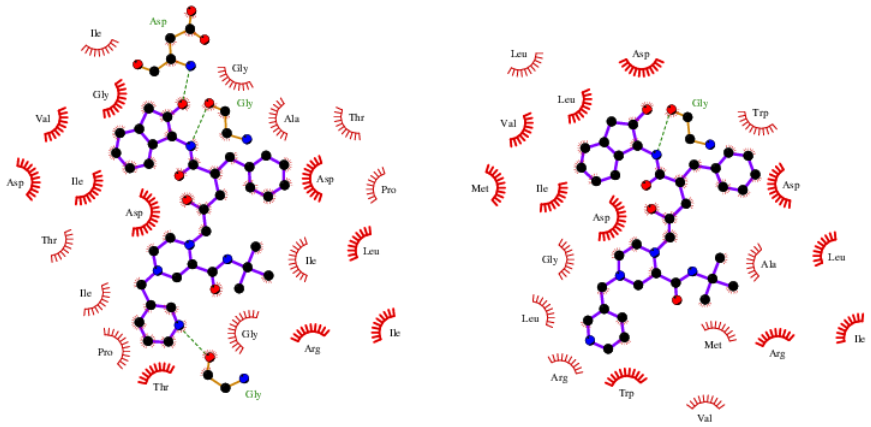


Figure 3.16 The interactions of indinavir (MK1) in its crystal structure in complex with the HIV-1 Protease from **1C6Y** [135] (left pane) and HTLV-1 Protease from **3WSJ** [137] (right pane).

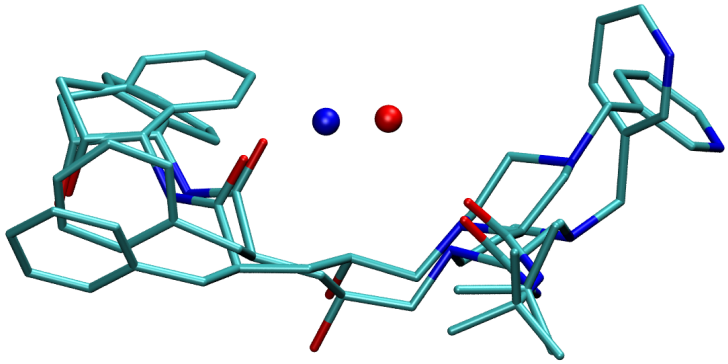


Figure 3.17 Overlay of indinavir from complexes with HTLV-1 and HIV-1 proteases showing the location of crystallized water in each *holo* structure as red/blue spheres.

Figure 3.18 shows the *apo* hydration of the HIV-1 (left-pane) and HTLV-1 (right-pane) binding sites predicted through MD. Contrary to 2TA in BRD4(1), *apo* HIV-1 and HTLV-1 proteases are both hydrated at the site of the ligand bound water. This water molecule located between the two flaps is conserved across all protease structures and has often been targetted in drug design[2]. With the structure of indinavir bound to a non-protease not availible, a direct assignment of the water molecule to protein or ligand in this case is not possible.

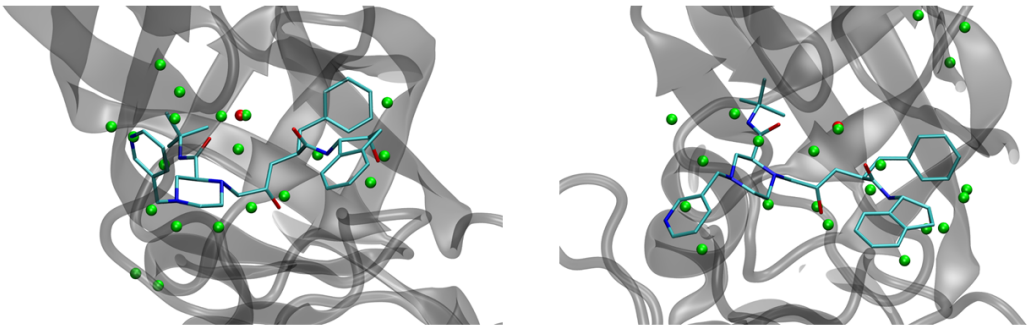


Figure 3.18 Overlay of the water bound to the indinavir with the locations of *apo* hydration sites from MD. Left Pane - **1C6Y**, Right Pane - **3WSJ**. The *apo* hydration sites are shown in green and the ligand bound water in red.

3.4.4.3 Cocaine

Cocaine is an amphiphilic tropane alkaloid [138, 139] extracted from the leaves of the *Erythroxylum Coca* plant and is the second largest drug of abuse worldwide [140]. It can cause potent cardiovascular and central nervous system stimulation and the effects of use include euphoria, agitation, delirium, increased heart-rate and loss of thermoregulation [141, 142]. Crystallographic structures of cocaine exist of it bound to dopamine transporter, ACh-binding protein and catalytic antibodies. Figure 3.19 shows a comparison of the binding site interactions of cocaine when bound to the dopamine transporter (**4XP4** [143]) and the GNC92H2 antibody (**1I7Z** [144]). Being two completely different family of proteins, the binding site interactions of cocaine in the two cases are completely different. Despite this, both

[2]See Section 1.2

structures have water molecules crystallized at the high-occupancy site predicted by MD (see Figure 3.20).

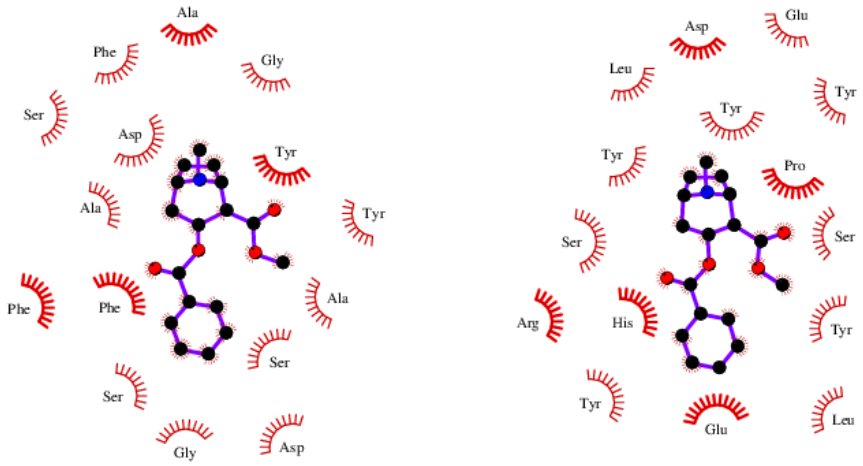


Figure 3.19 The interactions of cocaine in its crystal structure in complex with GNC92H2 antibody **1I7Z** (left pane) and drosophila dopamine transporter from **4XP4** (right pane).

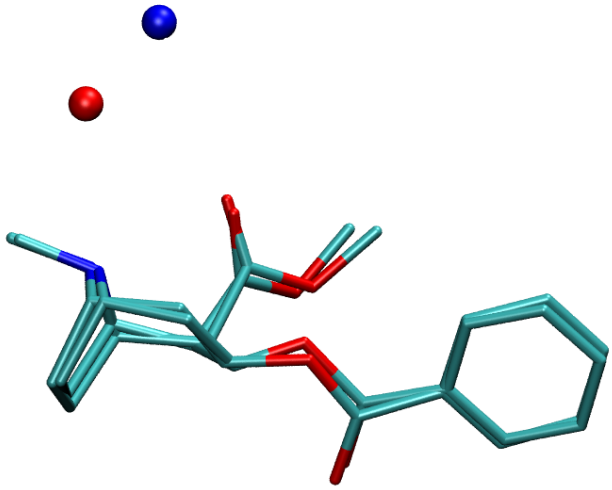


Figure 3.20 Overlay of cocaine from complexes with GNC92H2 antibody and drosophila dopamine transporter showing the location of crystallized water in each *holo*-structure as red/blue spheres.

Figure 3.21 shows the *apo* binding site hydration of GNC92H2 antibody (**1I7Z**) and drosophilia dopamine transporter (**4XP4**) overlaid with the site of ligand bound molecule. In the absence of cocaine bound, both *apo* proteins are not hydrated at the site of the ligand bound water.

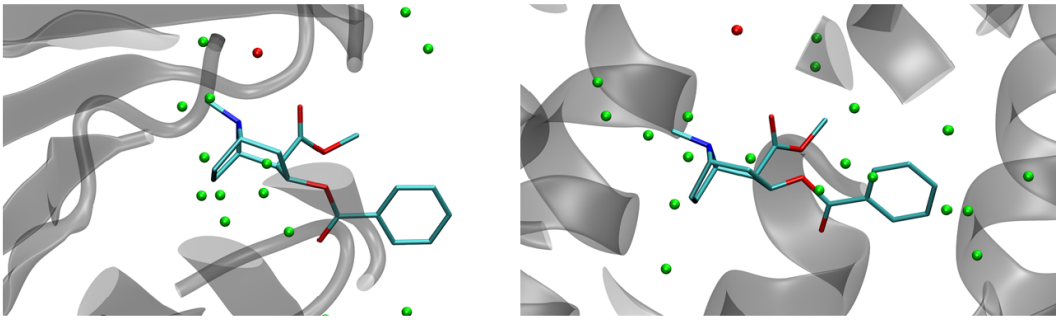


Figure 3.21 Overlay of the water bound to the cocaine with the locations of *apo* hydration sites from MD. Left Pane - **1I7Z**, Right Pane - **4XP4**. The *apo* hydration sites are shown in green and the ligand bound water in red.

3.4.4.4 Sucrose

Sucrose ($C_{12}H_{22}O_{11}$) is a naturally occurring carbohydrate found in many plants and plant parts. It is a disaccharide sugar and is a combination of two monosaccharides glucose and fructose [145]. A search of the Protein Data Bank reveals 222 *holo*-structures with sucrose bound either as a ligand or a co-factor. Figure 3.22 shows a comparison of the binding site interactions of sucrose when bound to the human splicing factor dim2 (**3GIX** [146]) and beta-lactamase (**1YLJ** [147]). In both cases, water molecules are crystallized at the high occupancy site predicted by MD (see Figure 3.23). The relevance of ligand-bound water is most evident in **3GIX** where a water molecule is present despite sucrose notbeing crystallized within an enclosed pocket.

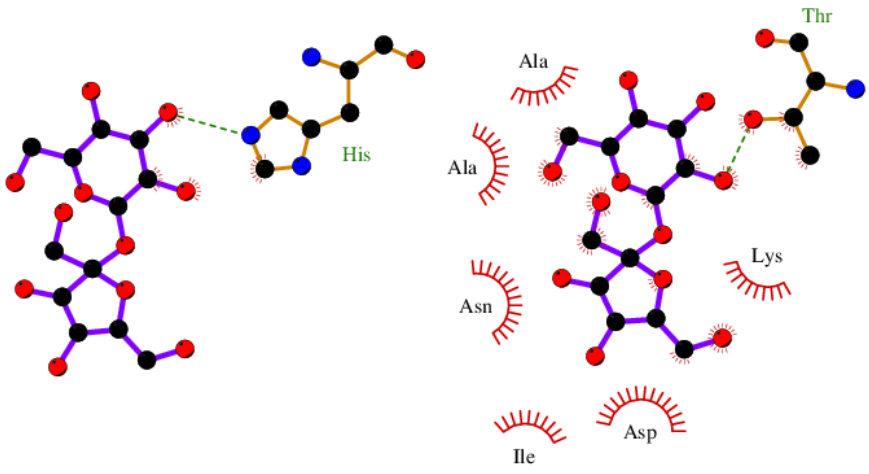


Figure 3.22 The interactions of Sucrose in its crystal structure in complex with the human splicing factor **3GIX** (left pane) and beta-lactamase from **1YLJ** (right pane).

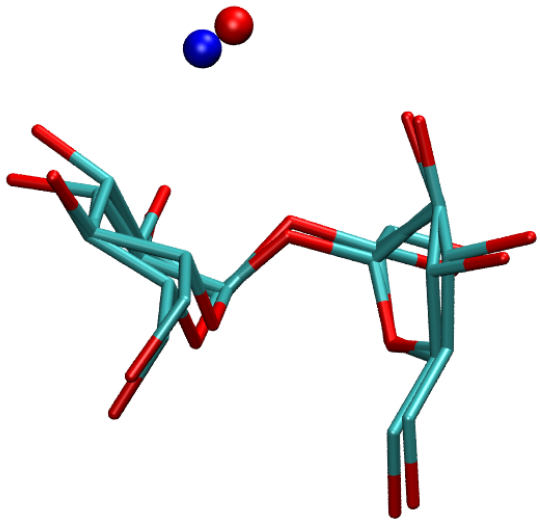


Figure 3.23 Overlay of sucrose from complexes with the human splicing factor and beta lactamase showing the location of crystallized water in each *holo*-structure as red/blue spheres.

Figure 3.24 shows the binding site *apo* hydration of human splicing factor dim2 (**3GIX**) and beta-lactamse (**1YLJ**) overlaid with the site of ligand bound molecule. In the absence of sucrose bound, both *apo* proteins are not hydrated at the site of the ligand bound water.

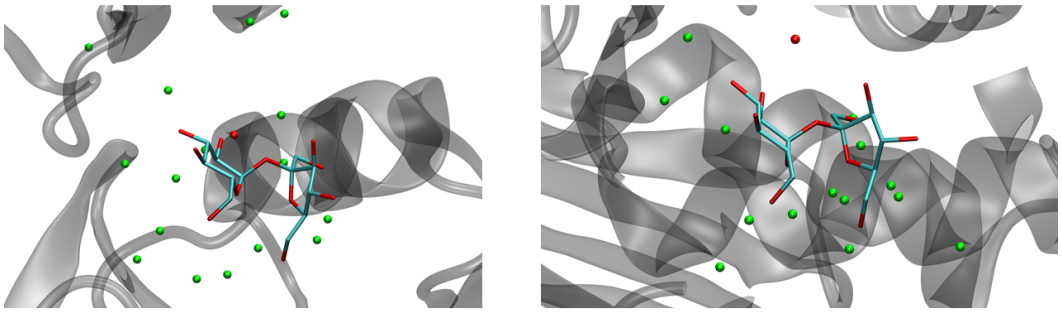


Figure 3.24 Overlay of the water bound to the sucrose with the locations of *apo* hydration sites from MD. Left Pane - **3GIX**, Right Pane - **1YLJ**. The *apo* hydration sites are shown in green and the ligand bound water in red.

Thus, the four ligands TG101348, indinavir, cocaine and sucrose have increasingly diverse binding site interactions. However, all four ligands have water molecules crystallized at the high occupancy site predicted by MD. This suggests that these waters are to be considered an intrinsic part of the ligand in their binding studies.

3.5 Conclusion and Further Work

Molecular Dynamics simulation of ligands and the identification of their hydration sites suggests a caveat to the traditional view that conservation/displacement of water molecules is completely ligand independent. While protein interactions do seem to play the deterministic role, the locations of all bridging water molecules overlapping with ligand hydration sites suggests that favourable interactions with the ligand is necessary.

More interestingly, hydration scoring also identifies four major drugs/ligands with high occupancy water-sites - TG101348, indinavir, cocaine, sucrose. These ligands representing $\sim$1% of the dataset have water molecules conserved across different *holo* structures. This suggests that these waters are to be considered an intrinsic part of the ligand in their binding studies. While only five structures were identified in this work, a similar proportion within the millions of structures typically analyzed by drug-screening protocols [35] might represent a sizeable number of molecules whose ligand-bound waters can mediate interactions with the protein. Simulations of a larger dataset of ligands can help in the identification of more such molecules and also acccurately validate the scoring metric used here.

Chapter 4

Predicting Bridging Waters

4.1 Aims and Summary

In this chapter, I introduce WaterDock 2.0, an adaptation of the WaterDock prediction pipeline previously developed in the lab by Dr. Greg Ross. Bridging water molecules play an important role in the binding of ligands and using them in drug-design first requires an accurate prediction of their location. The results of Chapter 3 show that bridging water molecules are a subset of ligand hydration sites. The new pipeline incorporates this into WaterDock to allow comparable accuracy in predictions albeit with a markedly reduced number of false-positives. The methodology is scripted in Python and implemented as a PyMOL plugin for wider use and adoption.

4.2 Introduction

As discussed in Chapter 1, water molecules are a key structural feature of protein-ligand complexes. Hence, bridging water molecules are increasingly being included in a variety of computational binding studies [148–150, 51]. In Quantitative structure-activity relationship (QSAR) modelling, Hussain et al. [151] found that the incorporation of explicit water molecules in the binding-site of actin enabled better calculations for ligands with the formamide moiety. Similarly, Taha et al. [152] incorporated water molecules in the 3D contact analysis search for inhibitors of candida N-myristoyl transferase (NMT) and glycogen phosphorylase (GP). A variety of groups have also used crystallographic water molecules in molecular docking screens. Huang et al. [126] used explicit water sites in 24 proteins to improve their docking enrichment factors and reduce false positives. Verdonk et al. [36] used crystallographic waters in the Astex Diverse set [55] to improve docking performance by upto 20%. The use of water molecules in docking studies has become so ubiquitous that programs like AutoDock4 [63], Gold [36], Rosetta [64], Glide [153] and FlexX [53] all offer options to include explicit waters.

As discussed in Section 1.1, a number of '*displaceability*' protocols have been developed to predict the location of waters in *holo*-structures. While these protocols are both computational [19, 23, 24] and knowledge-based [27, 28] in nature, their utilization requires a prior knowledge of *apo* hydration. This is often unknown if the crystallographic resolution is low [58] or if the structures were resolved through NMR [60]. Structures developed through homology modelling also have no details of hydration sites. While Molecular Dynamics (MD) or Monte Carlo (MC) simulation followed can identify peaks in 3D corresponding to the location of water-sites [67], the use of water molecules in drug screening requires an expeditious prediction of their locations. AcquaAlta [74] and WaterDock [29] are two such prediction protocols that use semi-empirical and knowledge based methods to swiftly predict probable water locations in binding sites.

AcquaAlta utilizes crystallographic information from the Cambridge Structural Database (CSD) [75] to predict probable water-sites. The preferential orientations of water across functional groups are obtained from a search of the CSD. The intersection of these favourable geometries between the ligand and protein are then used to predict bridging water-sites. When tested on a dataset of 14 high-resolution crystallographic structures of the Oligopeptide Binding Protein (OppA), AcquaAlta could predict 66% of the bridging waters.

WaterDock, a pipeline devised by Ross et al. [29] uses the freely available AutoDock Vina [72] tool to predict the location of binding site waters. Figure 4.1 summarizes the prediction pipeline of this protocol. Water molecules are initially treated as ligands and are docked thrice onto the binding site of the protein. With AutoDock Vina able to predict upto 20 configurations per run, the initial docking results in a maximum of 60 probable water co-ordinates. A Vina score cutoff of $\leqslant -0.6\, kcal/mol$ is then applied to the co-ordinates to remove water-sites that are energetically unfavourable. The ensemble of water co-ordinates are then clustered twice using the single linkage method with distance cut-offs of 0.5 Å and 1.6 Å.

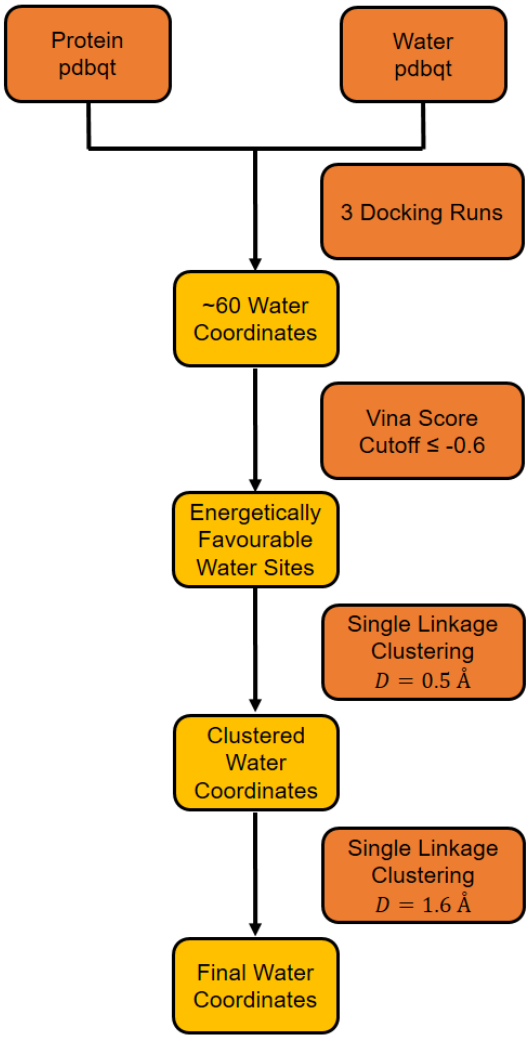


Figure 4.1 Schematic illustration of the WaterDock prediction pipeline.

The cut-off Vina scores and the clustering distance radii were empirically calculated after training on 15 crystallographic structures of 9 proteins. The protocol was then tested against consensus water-sites in 7 proteins that had been resolved more than once. Table 4.1 lists the PDB codes of the seven complexes used to identify the consensus water-sites. Additionally, WaterDock was also tested on the set of OppA structures used in the validation of the AcquaAlta algorithm.

Table 4.1 The structures used to identify consensus water-sites in the Water-Dock algorithm.

Protein	PDB codes	Ligand
HIV-Protease	3FXS,1HPX,2ZYE	KNI-272
Ribonuclease A	1KF5, 1FS3,5RSA	None
GluR2	1FTM, 1MY2	AMPA
Trypsin	1S0Q, 1UTQ, 1TPO	None
Concanavalin	1NLS, 1GKB, 1JBC, 1QNY	None
Gluthathione S-transferease	1K3Y, 1K3L	S-hexyl gluthathione
Carbonic Anhydrase	3KS3, 3MWO, 1ILI	None

WaterDock was able to predict upto 88% of the consensus water-sites in the seven crystal structures and 97% of the water structures in the OppA Dataset. However, it over-estimates the number of discrete water-sites within the binding pocket. In the tests against consensus water-sites from the seven crystal structures, the overall false-positive rate was upto 24%. When applied to the AcquaAlta OppA dataset, the false positive rate was on average 1-2 water sites per structure. Additionally, the clustering procedure of WaterDock is optimized for a box of side 15 Å. Hence, in the case of larger ligands such as **2TA** [134, 129] described in section 3.4.4, the protocol must be applied more than once to boxes of side 15 Å each.

In conclusion, the algorithms that predict bridging water sites in *holo* structures either necessitate a prior knowledge of *apo* hydration or have low accuracy/high false positive rates. The results of Chapter 3 show that bridging water molecules are a subset of ligand hydration sites. WaterDock 2.0 aims to combine this result with the protocol laid out by Ross et al. [29] to reduce prediction false positives and to allow the protocol to be applied directly to a wider set of *holo* structures.

4.3 WaterDock 2.0 Pipeline

The results of Chapter 3 showed that bridging water molecules are a subset of ligand hydration sites. However, the strength of ligand-water interactions in these sites rarely affected the potential to be conserved in *holo* structures. WaterDock 2.0 was thus envisaged to combine ligand hydration with the scoring of water-protein interactions to better predict bridging water sites. Figure 4.2 describes the pipeline of the new prediction algorithm.

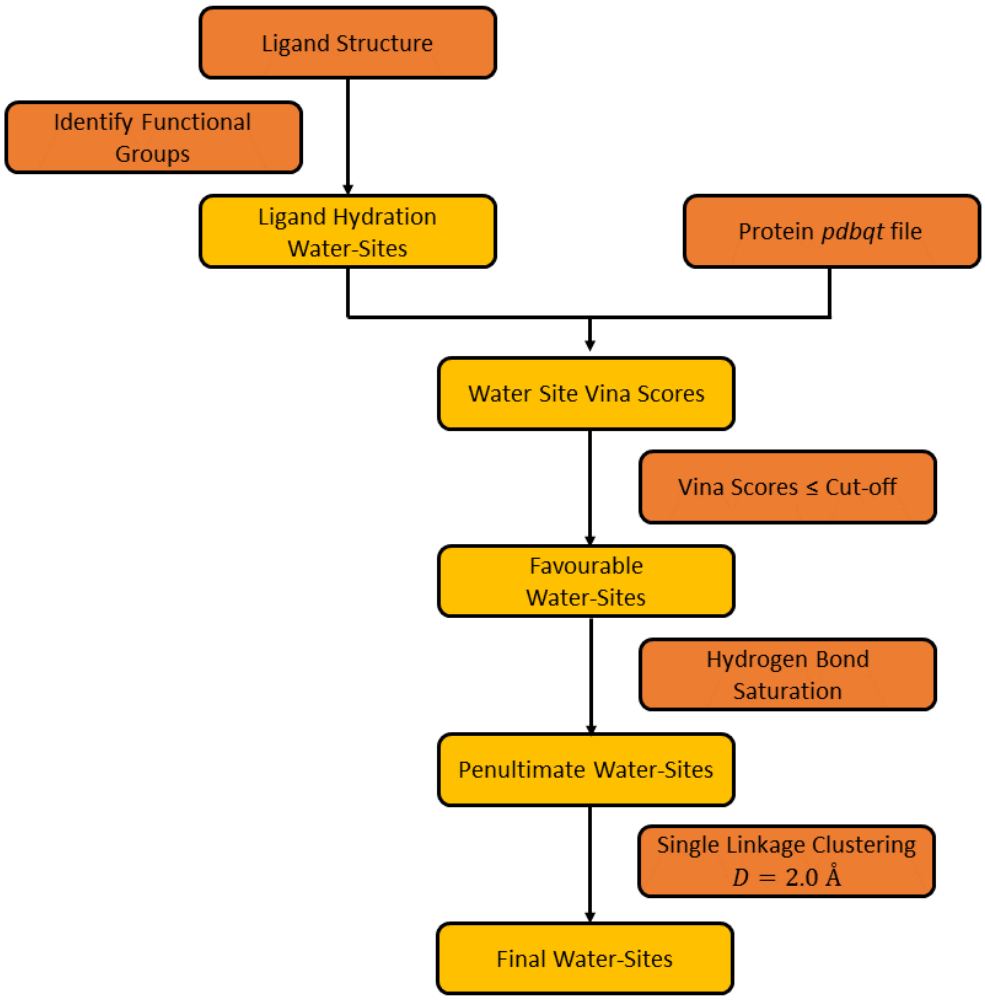


Figure 4.2 Schematic illustration of the WaterDock 2.0 bridging water prediction pipeline.

To speed-up the generation of the hydration shell, a semi-empirical approach initially utilizes functional group orientations to map water-sites. Protein-water

interactions for each of the probable water-sites are then individually scored using AutoDock Vina [72]. Subsequently, unfavourable water-sites with a Vina score lower than an empirically calculated cut-off are discarded. As the protocol uses water-protein interactions as the only scoring metric, false-positive results are reduced through the use of a hydrogen bond saturation limit. This limit dictates the maximum number of plausible bridging water-sites from the hydration shell of each functional group. Thus, if more water-sites from a motif's hydration shell have favourable Vina scores, only the highest scoring sites among them are selected. Table 4.2 lists the H-bond saturation limit enforced on each of the functional groups used in this work. The saturation limits were selected based on the number of VSEPR lone pairs or bound hydrogens in the oxygen/nitrogen atoms.

Table 4.2 The H-bond saturation limit enforced on the various functional groups used in the WaterDock 2.0 pipeline.

Motif	H-bond limit
Carbonyl	2
Carboxyl	2
Cyano	1
Imine	1
Nitro	2
Amines	No. of bound H
Sulfonyl	2
Phosphoryl	2
Hydroxyl	3
Ether	2
Halogen	1

Finally, water-sites within a diameter of 2.0 Å of each-other are clustered to avoid the hydration sites subtended by multiple ligand functional groups to be predicted as multiple bridging waters.

4.4 Methods

4.4.1 Identifying Functional Group Hydration

4.4.1.1 Methodology

Ligand structures were initially checked for their correct protonation states before being solvated in a TIP3P [82] rectangular box of water with a minimum distance of 14 Å from each edge. The partial charges on the atoms were calculated according to the AM1-BCC [102, 103] method using *Antechamber* [108] and the solvated box was neutralized by the addition of Na^+/Cl^- ions as paramterized by Joung et al. [117]. All boxes were simulated for 30 ns each with a time-step of 2 fs. The temperature was maintained at 300 K and the pressure at 1 bar using the V-rescale thermostat and the Parinello-Rahman [118] barostat respectively. Co-ordinates were saved every 0.6 ps resulting in 50000 trajectory frames. During the course of the simulation, the conformation of the ligands were maintained using positional restraints of 100 $kJ/mol\,nm^2$ on all non-hydrogen ligand atoms. The total hydration distributions around all the ligands were then calculated according to the method described in Section 3.2.

Figure 4.3 shows the pipeline used to identify polar functional groups within the ligand. Firstly, the connectivity of the atoms are determined using the same method as one the used in *Antechamber* and described in Section 2.2.3. For oxygen and nitrogen atoms, their functional groups are determined by calculating the number atoms they are bound to. Oxygen atoms '*singly bound*' to S, P and N atoms are classified as sulfonyl, phosphoryl and nitro functional groups

respectively. Oxygen atoms '*singly bound*' to carbon atoms are classified by checking the connectivity of the C atom. If an additional oxygen atom is not bound to the carbon, the oxygen is classified as a carbonyl. However, the presence of an additional oxygen atom means its connectivity is also validated to allow discrimination between ester (carbonyl O + ether O) and carboxylic acid functional groups. Oxygen atoms '*doubly bound*' are classified as ether or hydroxyl functional groups based on the number of Hydrogen atoms bound to it.

Nitrogen atoms '*singly*' or '*doubly bound*' are classified as cyano and imine functional groups respectively. Nitrogen atoms '*triply bound*' are classified as nitro or amine functional groups depending on the number of bound oxygen atoms. Finally quartenary nitrogen atoms are classified as amines.

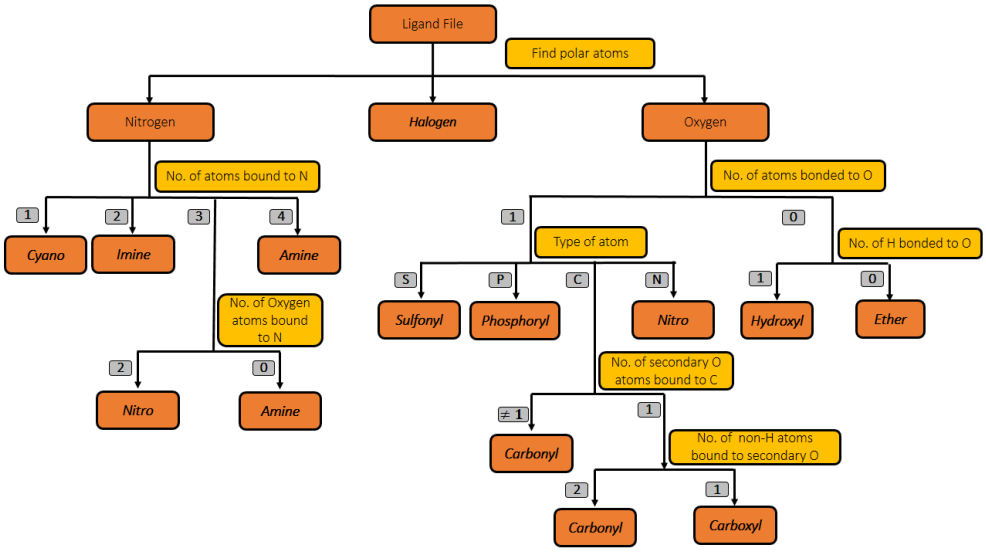


Figure 4.3 Schematic illustration of the pipeline used to identify functional groups in ligand molecules.

After identification of functional groups in each ligand, all hydration sites were then 'assigned' to the nearest functional group. The 'assigned' hydration sites of each functional group were then totalled across all ligands in the dataset. This allows a comprehensive picture of probable functional group hydration distribution across different atomic surroundings.

4.4.1.2 Dataset

The hydration of functional groups was calculated using the previously simulated CSAR-2012 dataset [66]. With water molecules not reliably resolved in the 242 crystallographic structures, the dataset has limited use in the development of prediction algorithms. However, the *holo* structures contain ligands resolved with no conformational ambiguity. This structural accuracy and the large number of molecules in the dataset make it ideal for compiling the hydration of specific functional groups across a range of chemical environments.

4.4.2 Identifying Energetic Cut-off

4.4.2.1 Methodology

Hydrogen atoms were added to the ligand structures using the Reduce program [115] and their protonation states were visually verified. The functional groups of the ligands were identified and then used to 'map' the hydration shell of the molecule. The python script accompanying AutoDock [52] was then used to generate *pdbqt* files of the protein structures. The protein-water interactions of each of the water molecules in the ligand's hydration shell was scored using AutoDock Vina. To avoid Vina's search algorithm from displacing the co-ordinates of the water molecules, each water-site in the ligand hydration shell was independently docked onto the protein with a miniscule box-size of 0.5 Å centred on its co-ordinates. While docking, the '*num_ modes*' option of Vina that controls the number of output configurations was set to 1 to allow generation of only the most favourable interaction score.

For each of the water-sites, its distance to the nearest crystallographic water was calculated and considered '*conserved*' if the distance was lesser than 2.0 Å.

4.4.2.2 Dataset

The energetic cut-off for the prediction algorithm was calculated using the Astex diverse dataset of 85 *holo* structures [55]. While all relevant bridging water molecules are not resolved within the dataset, the resolved sites can be considered accurate. Docking studies by Hartshorn et al. [55] using GOLD showed that the crystallized water molecules can be used to improve docking accuracy to greater than 90%. Thus, this validation set is chosen to find the energetic cut-off score that allows the prediction of the maximum number of water-sites. However, with not all bridging waters crystallized, a fairly high false-positive rate is envisaged in this training set.

4.4.3 Validation

The co-ordinates of water molecules are notoriously difficult to accurately estimate. During the development of WaterDock, Ross et al. [29] attempted to circumvent this by overlaying multiple independently crystallized structures and only considering those sites resolved at least twice. However, the original dataset consisted primarily of *apo* structures and only had three *holo* structures. The dataset to validate WaterDock 2.0 was compiled to include those initially used by Ross et al. and extended to include other high-resolution *holo* structures crystallized multiple times. The ligands in each structure were overlaid and the water-coordinates within 3.4 Å of both ligand and protein present more than once were considered as consensus sites. Table 4.3 lists the details of the validation dataset and the number of consensus water-sites in each case.

Table 4.3 The dataset of structures used validate the ligand-directed Waterdock 2.0 algorithm.

Protein	PDB codes	Ligand	Resolution (Å)	Consensus Sites
HIV-1	3FXS,1HPX,2ZYE	KNI-272	0.93, 2.0, 1.9	4
GluR2	1FTM, 1MY2	AMPA	1.7, 1.8	3
Trypsin	2AH4, 3RXJ	GBS	1.13, 1.7	4
GST	1K3Y, 1K3L	GTX	1.3, 1.5	6
HSP90	2BRC, 2BT0	CT5	1.6, 1.9	2
PIM1	1XWS, 2BIK	RBT205	1.8, 1.8	1

To allow a direct comparison with AcquaAlta and the original WaterDock protocols, the 14 OppA structures used by both were also analyzed. In OppA, water molecules have been postulated to play a role of flexible mediators with their rearrangement allowing a wider array of ligands to bind. The ligands in the 14 structures are tripeptides of the form K-X-K and the number of bridging water molecules vary from 6 to upto 10. Table 4.4 lists the details of the OppA structures and the number of water molecules in each.

Table 4.4 The OppA dataset of structures used to allow comparison of new prediction protocol to AcquaAlta and the original WaterDock methodologies.

PDB codes	Ligand	Resolution (Å)	Water-Sites
1JET	KAK	1.2	7
1JEU	KEK	1.25	9
1JEV	KWK	1.30	6
1B4Z	KDK	1.75	10
1B5I	KNK	1.9	7
1B32	KMK	1.75	7
1B3F	KHK	1.8	7
1B46	KPK	1.8	6
1B51	KSK	1.8	9
1B58	KYK	1.8	7
1B5J	KQK	1.8	10
1B9J	KLK	1.8	6
1QKA	KRK	1.8	6
1QKB	KVK	1.8	6

4.5 Results

4.5.1 Functional Group Hydration

Figure 4.4 shows the total hydration structure calculated from MD around five polar functional groups - carbonyl, carboxyl, ether, phosphoryl and imine. The orientation of water-sites for each group is also analyzed as a distribution of ϕ and θ - the angles made by the ligand atom - O_w vector with an imaginary **X** and **Z** axes (see Figure 4.5). While IsoStar [154] can provide similar information from the CSD, its unavailibility necessitated the hydration of functional groups to be computationally compiled from MD.

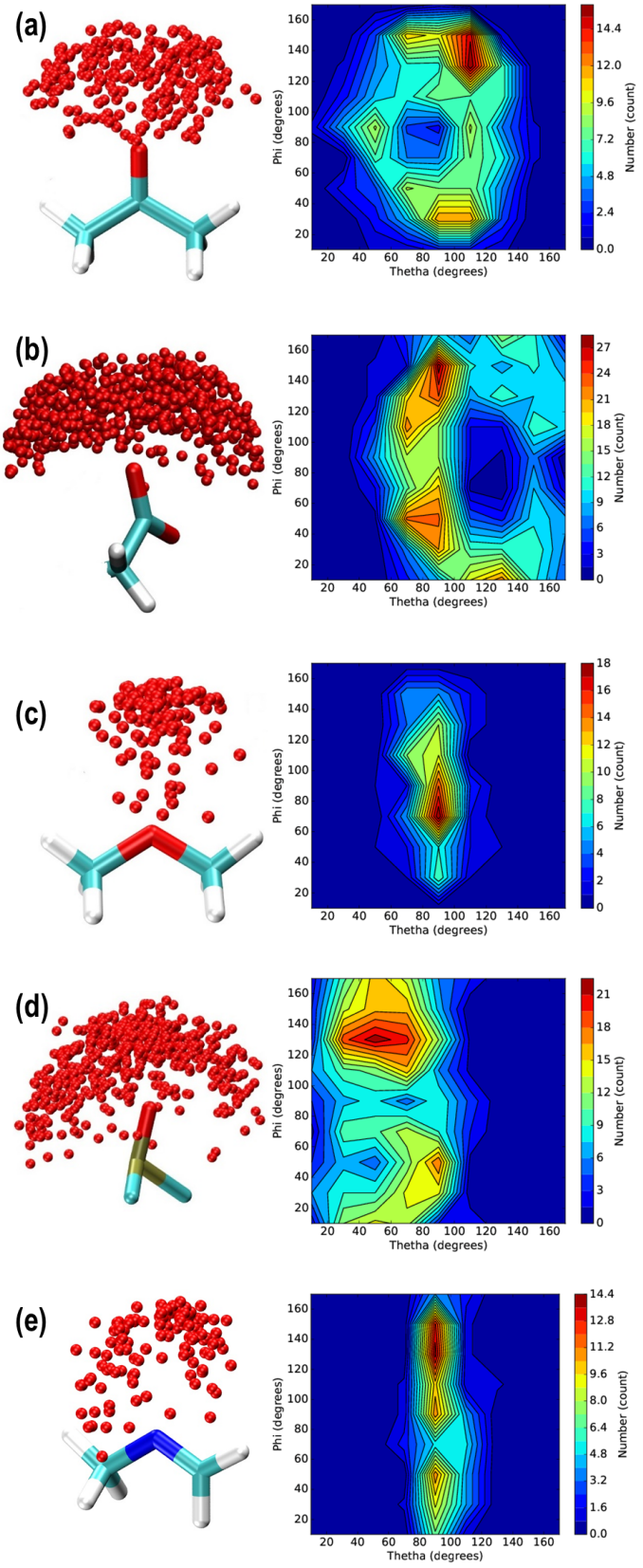


Figure 4.4 The hydration structure of five ligand polar functional groups from MD **(a)** carbonyl **(b)** carboxyl **(c)** ether **(d)** phosphoryl **(e)** imine. The distribution around the polar groups are plotted in the right pane as a 2D histogram of the angles θ and ϕ. The results are obtained by overlaying the specific functional group's closest hydration sites from the MD simulations of 242 ligands from the CSAR-2012 dataset.

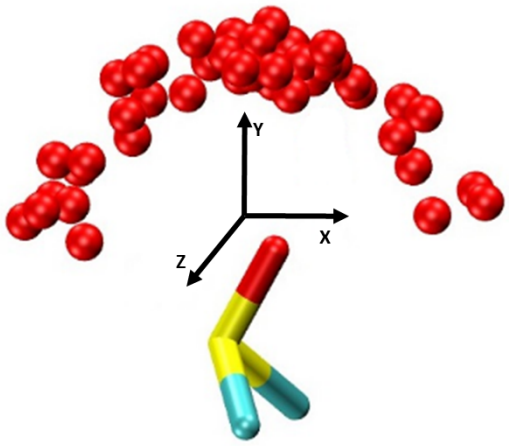


Figure 4.5 Schematic illustration of the imaginary X, Y and Z axes placed around the functional groups. The angle made by the ligand atom - O_w vector with the **X** and **Z** axes are ϕ and θ respectively.

Hydration sites around carbonyl oxygens are distributed in an uniform 'oval' shape with θ ranging from 40° to 140° and ϕ ranging from 20° to 160°. The distribution around carboxyl oxygens is also 'oval' albeit with a larger range of angles with θ spanning from 60° to 160° and ϕ spanning from 0° to 180°. This increased range of binding might probably be attributed to a combination of the larger solvent accessibility of carboxyl oxygens and the greater negative charge on the oxygen atoms. Additionally, θ deviates from a mean value of 90° seen in carbonyl oxygens to $\sim$ 110° probably due to hydration sites also interacting with the other oxygen atom of the carboxylic motif. In phosphoryl oxygens, the distribution of hydration sites is similar to that seen in carboxylic acids with a mean θ value again deviating from 90°. This deviation might probably be caused by a similar reason with phosphoryl oxygen atoms normally occuring as pairs in ligand molecules. The hydration sites of both ether oxygens and imine nitrogens lie on a plane normal to that of the motif with θ not deviating from $\sim$ 90°.

The simulation results are used to model the hydration structure of the five functional groups. First a vector $\vec{v}$ normal to the XZ plane and modelled as $3.0 \times y_{axis}$ is placed on the polar atom. $\vec{v}$ is then rotated around the **X** and **Z** axes using the Euler-Rodriguez rotation formula to match the results of Figure

4.4. The Euler-Rodriguez vector rotation formula [155] provides $\vec{v}_{rot}$, the vector obtained by rotation of $\vec{v}$ around $\vec{k}$ by an angle of σ.

$$\vec{v}_{rot} = \vec{v}\cos(\sigma) + (k \times \vec{v})\sin(\sigma) + k(k \cdot \vec{v})(1 - \cos(\sigma)) \tag{4.1}$$

With the polar atom - O_w vector modelled with length 3.0 Å, σ was set as 40° to ensure the modelled water-sites are 2.0 Å apart. Figure 4.6 shows the hydration sites of the five functional groups 'modelled' to match the results of Figure 4.4.

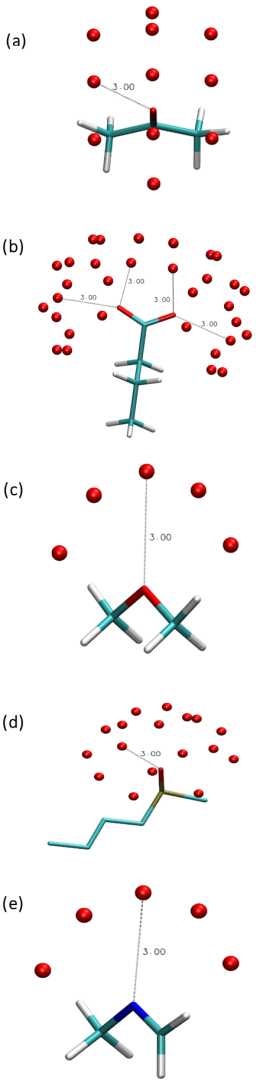


Figure 4.6 The modelled hydration structure of five ligand polar functional groups **(a)** carbonyl **(b)** carboxyl **(c)** ether **(d)** phosphoryl **(e)** imine.

Figure 4.7 shows the location of water-sites around different amines in both protonated and unprotonated states. In all cases, hydration sites preferentially bind to the H atom with limited direct binding to the N atom. Hence, amine hydration sites are modelled along the direction of the N-H vector with a distance of 3 Å from the nitrogen atom (Figure 4.8).

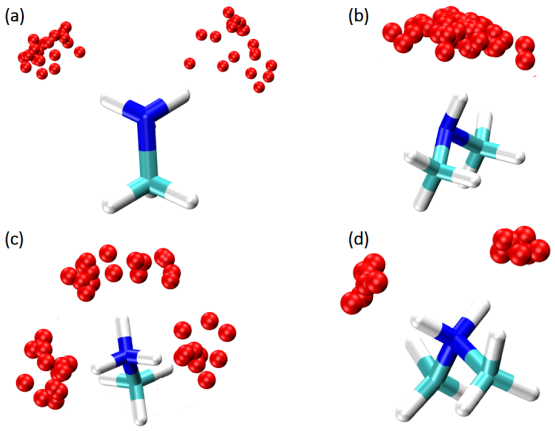


Figure 4.7 The location of water-sites around different amine groups from MD. **(a)** Unprotonated primary amine **(b)** Unprotonated secondary amine **(c)** Protonated Primary amine **(d)** Protonated secondary amine.

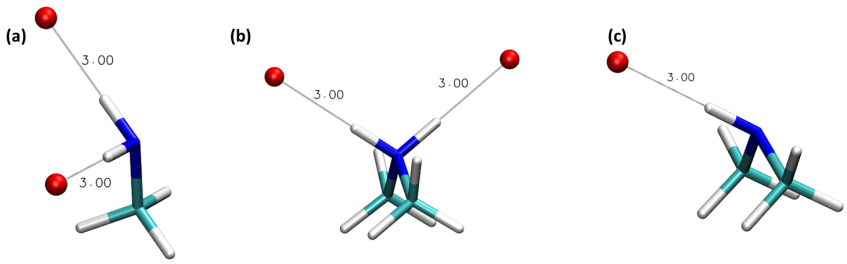


Figure 4.8 The modelled hydration structure of amine groups.

Figure 4.9 a shows the hydration structure of hydroxyl groups from simulation. Two distinct orientations of water are seen bound to the hydrogen and oxygen atoms. Therefore, the hydroxyl group is modelled as a combination of an 'ether' oxygen atom and an 'amine' hydrogen atom (Figure 4.9 b)

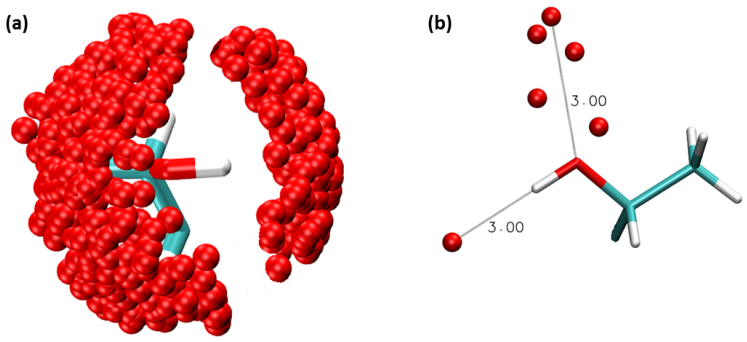


Figure 4.9 (a) The distribution of hydration sites around the hydroxyl functional group from simulation. **(b)** The modelled hydration structure of the hydroxyl functional group.

Of the 55 chloride functional groups in the dataset, none had a water-site distributed around it in MD simulation. On the other hand, Fluorine atoms were hydrated. However, with only four F atoms in the dataset, an accurate hydration distribution could not be determined. Similarly, insufficient ligands with 'nitro', 'sulphonyl' and 'cyano' functional groups were present in the dataset. The hydration of all four functional groups were thus modelled similar to that of the most extensively hydrated motif - carboxyl oxygen.

4.5.2 AutoDock-Vina Energetic Cut-off

Modelling the hydration sites based on functional groups predicted 2045 water-sites around the 85 ligands of the Astex Diverse Set. Figure 4.10 plots the results of the docking study on each of the 2045 'modelled' water-sites. It shows a scatter plot of the site's docking Vina-Score against the distance to the nearest crystallographic water. The previously established distance cut-off of 2.0 Å is also plotted as dotted vertical line.

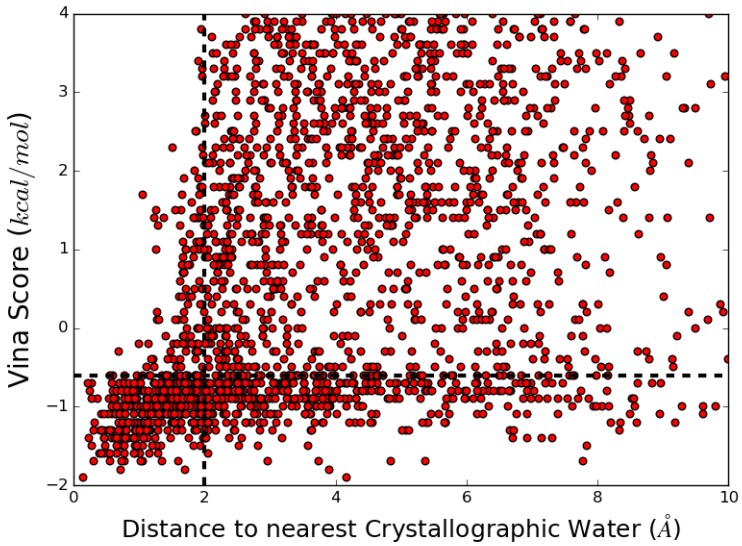


Figure 4.10 Scatter plot showing the water-site's Vina docking score against the distance to the nearest crystallographic water on application of the WaterDock 2.0 pipeline to the Astex Diverse Set.

The scatter plot reveals a correlation between the Vina binding score and the distance to a crystallographic water site. Ligand hydration sites with stronger binding scores are more likely to be '*conserved*' and vice-versa. However, for application to a Docking protocol, a cut-off score needs to be calculated (horizontal dotted line in Figure 4.10). Ligand hydration sites bound to the protein stronger than the cut-off are considered '*conserved*'. A decrease of this cut-off score would increase the number of '*conserved*' water-sites and thereby increase both the prediction and false-positive rates. Figure 4.11 plots the prediction and false-positive rates across a range of cut-off scores.

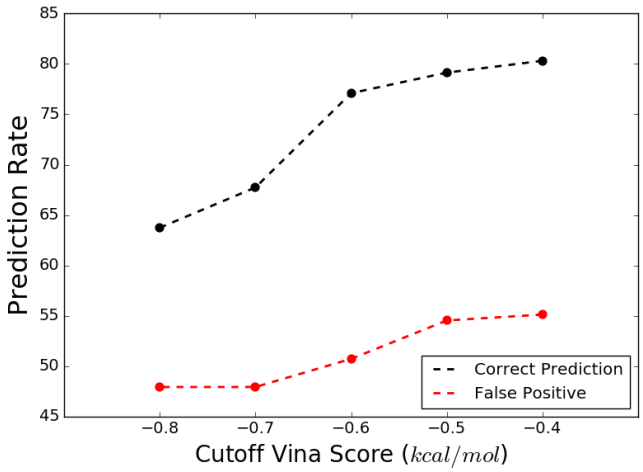


Figure 4.11 Plot showing the prediction and false-positive rates across a range of cut-off scores when the WaterDock 2.0 pipeline is applied to the Astex Diverse Set.

The statistical program **R** and its *rtree* package [156] were used to calculate the ideal cut-off score that maximizes prediction rate while minimizing false-positives. **R** calculated the ideal cut-off score to be -0.55 kcal/mol (shown as black horizontal line in Figure 4.10). Thus, all ligand hydration sites with a Vina score more positive than -0.55 kcal/mol are discarded as unfavourable.

4.5.3 Validation

Table 4.5 lists the results of the WaterDock 2.0 pipeline on the validation dataset of *holo* structures. A water-site was considered correctly predicted if it was within 2.0 Å from both water molecules used to identify a consensus site. The protocol is able to predict upto 84 % of the consensus water-sites in the protein structures while only predicting 1 non-consensus site.

Table 4.5 Individual Protein bridging water-site prediction results using the WaterDock 2.0 pipeline

Protein	Consensus Sites	Predicted Consensus Sites	Predicted False Positives
HIV-1	4	4	0
GluR2	3	3	0
Trypsin	4	3	0
GST	6	4	0
HSP90	2	2	0
PIM1	1	1	1
TOTAL	20	17	1

Figure 4.12 shows the location of the crystallographic and predicted water-sites for the six *holo* structures. The crystallographic waters from the two structures are shown in blue and red while the predicted waters are shown in green. Additionally, the correctly identified, missed and false-positive sites are highlighted in black, red and blue respectively. The incorporation of ligand conformation and hydration to the WaterDock protocol dramatically reduced the false-positive rate from $\sim 24\%$ to 5% . Moreover, the 'false-positive' site predicted in PIM1 has a crystallographic water in one structure. This raises the possibility that the second PIM1 structure missed the bridging water and that the true false-positive rate is actually lower. This reduction is possibly due to the H-bond saturation limit enforced on ligand functional groups. This limit reduces the probability of multiple water molecules predicted within a hydrophilic pocket of the protein. Additionally, unlike the original WaterDock protocol, the new pipeline also subtly includes water-water interactions due to the ligand hydration sites being mapped from MD simulations.

However, the new pipeline sacrifices the AutoDock Vina's search algorithm for semi-empirically generated probable sites. This results in a slightly greater mean error of 1.24 Å compared to the original mean error of 0.83 Å. Of the 17

water molecules correctly identified, 4 were greater than 1.5 Å from from either crystallographic water.

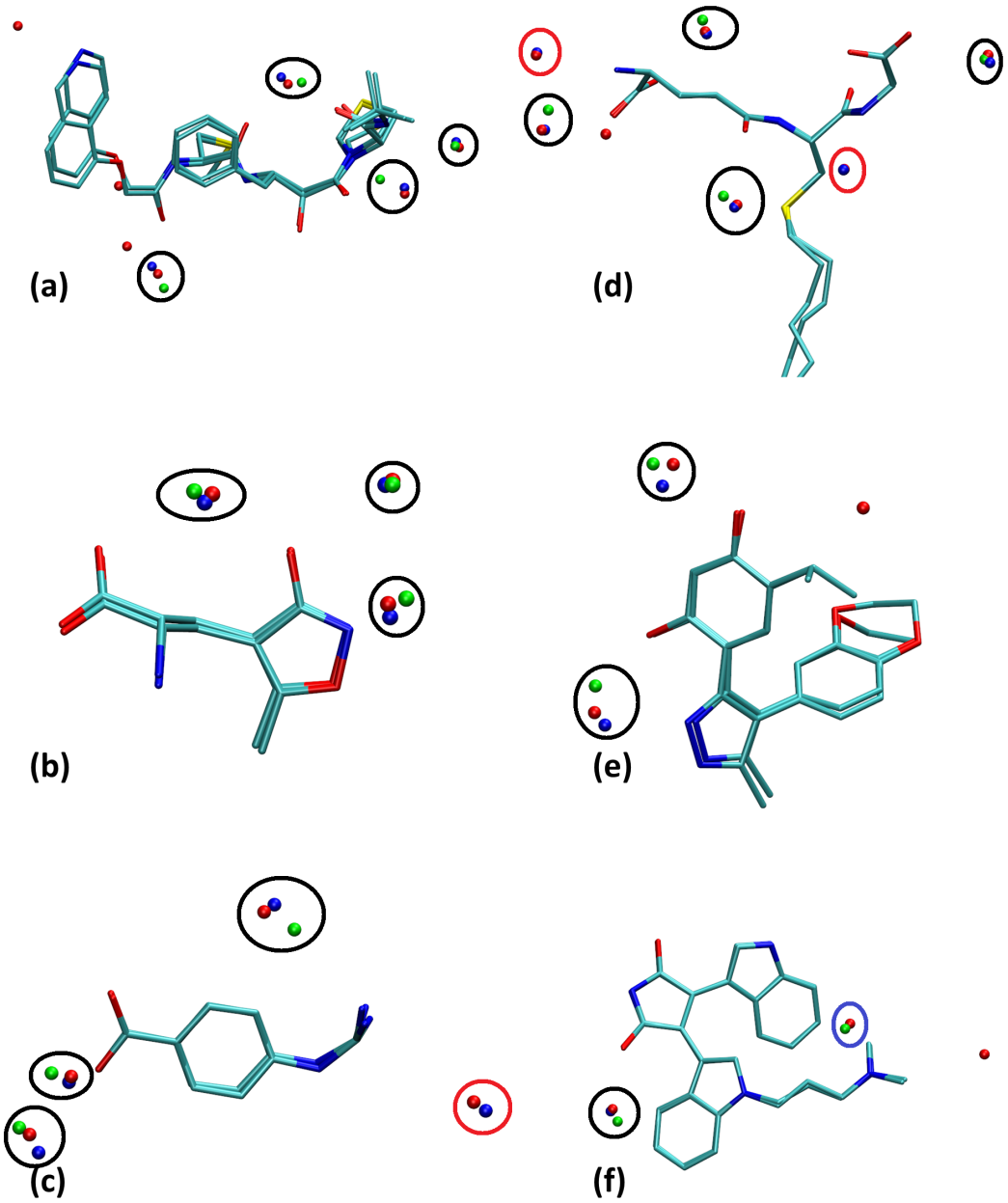


Figure 4.12 The location of the crystallographic and predicted water sites for the six protein structures. The crystallographic waters from the two structures are shown in blue and red while the predicted waters are shown in green. Additionally, the correctly identified, missed and false-positive sites are highlighted in black, red and blue respectively. The proteins are **(a)** HIV-1, **(b)** GlurR2, **(c)** Trypsin, **(d)** GST, **(e)** HSP90, **(f)** PIM1.

Table 4.6 lists the results of the OppA dataset used to further validate the protocol and offers a comparison to WaterDock and AcquaAlta. AcquaAlta was validated using a cut-off distance of 1.4 Å and could predict 66% of the 103

crystallographic waters. At the same maximum error WaterDock could predict 87% of bridging waters. The greater maximum error in the new protocol is apparent in the results reported in Table 4.6. At a cut-off distance of 1.4 Å, only 76 water molecules (74%) could be predicted. While inferior to the original protocol at this cut-off distance, the new pipeline performed much much better than AcquaAlta.

At a maximum error of 2.0 Å, WaterDock could predict 97% of the water molecules with a false-positive rate of $\sim$ 1-2 sites per structure. At this cut-off, the performance of the new pipeline was comparable to the original WaterDock protocol with a prediction rate of 93% . However, only six false-positive sites were predicted. The poor performance of the new pipeline at an error limit of 1.4 Å is expected as the probable ligand hydration sites are mapped with a spacing of 2.0 Å. The significantly improved false positive rate can possibly be attributed to the presence of carboxylic acid groups in all ligands of the dataset. This large functional group binds in a hydrophilic protein pocket within which WaterDock predicts multiple water molecules. However, the hydrogen bond saturation limit enforced in the new pipeline reduces this.

Table 4.6 The results of the OppA dataset of structures used to allow comparison of new prediction protocol to AcquaAlta and the original WaterDock methodologies.

Structure	Waters	Predictions (1.4 Å)	Predictions (2.0 Å)	False-Positives
1JET	7	6	6	1
1JEU	9	7	8	0
1JEV	6	4	6	1
1B4Z	10	7	8	0
1B5I	7	5	7	0
1B32	7	4	7	1
1B3F	7	6	7	0
1B46	6	4	6	1
1B51	9	7	8	0
1B58	7	5	6	0
1B5J	10	8	8	0
1B9J	6	5	6	1
1QKA	6	4	6	0
1QKB	6	5	6	1
TOTAL	103	76	95	6

In conclusion, the inclusion of ligand conformation and its associated hydration sites into the WaterDock pipeline allows a comparative accuracy in prediction with a markedly reduced number of false-postives. However, a trade-off is a slight increase in the mean error of the predictions.

4.6 PyMOL Plugin

PyMOL [123] is an extensively used open-source molecular visualization software. The **Py** portion of the software's name refers to the fact that it is extensible through the Python Programming language. These extensions to PyMOL are

mainly through Plugins - external modules that add functionality. PyMOL plugins have been developed for a range of modelling programs including Gromacs [157], Vina [72] and Amber [109]. Considering PyMOL's ubiquity and it's GUI improving ease of use, a plugin for the WaterDock 2.0 pipeline will probably enhance its appeal and adoption in the scientific community. The python *Tkinter*, *scipy* and *MDAnalysis* [119] libraries were used to develop a PyMOL plugin.

Figures 4.13 and 4.14 show snapshots of the plugin installed within PyMOL. The '*Apo* WaterDock' option is an implementation of the original WaterDock [29] previously scripted in R. The '*Holo* WaterDock' option is the WaterDock 2.0 pipeline developed. The plugins are currently available at *https://github.com/akshay-sridhar/WaterDock-2.0.git* .

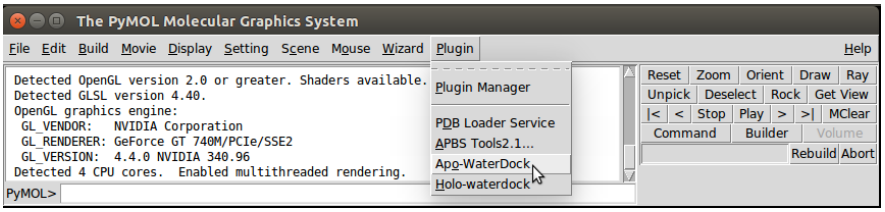


Figure 4.13 Snapshot of the PyMOL plugins developed for WaterDock pipelines.

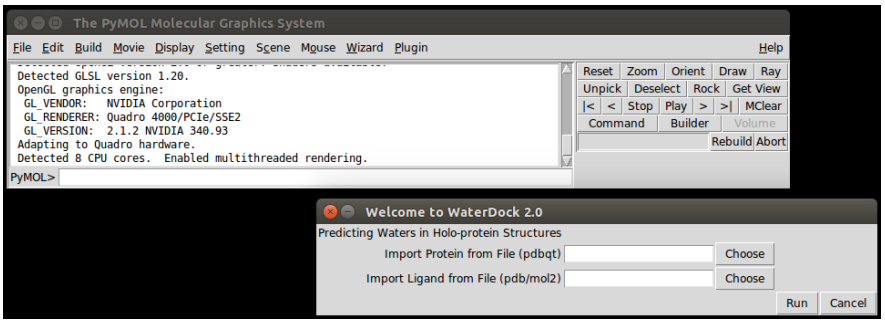


Figure 4.14 Snapshot of the PyMOL plugins developed for WaterDock pipelines.

4.7 Conclusion and Further Work

WaterDock 2.0 was envisaged with a view of combining ligand hydration with AutoDock Vina's scoring to predict water molecules within the binding site of *holo* structures. While the validation of any prediction protocol is inherently

difficult considering the inaccuracies associated with crystallographic waters, the new pipeline significantly reduced the number of false-positives while matching the the prediction accuracy of WaterDock. Additionally, the speed reduction caused by repeated independent docking runs is minimal as it is compensated for by a significant reduction in the search space of each run.

The swiftness of the protocol and its sensitivity to ligand conformation make it ideal for combination with drug screens. WaterDock 2.0 could be combined with docking poses to predict bridging water-sites which in-turn can be used to develop a 'hydrated' docking score. This methodology necessitates no prior knowledge of binding pocket hydration or docked conformations. While theoretically alluring, this 'hydrated' docking methodology requires further work and similar attempts by various groups have shown little performance improvement.

Chapter 5

An in-depth solvation analysis

5.1 Aims and Summary

Neutron Diffraction is the experimental method of choice to study small-molecule solvation. Empirical Potential Structure Refinement (EPSR) is a reverse Monte Carlo methodology used to fit atomic conformation to neutron diffraction data. In this chapter, I compare the solvation structures predicted by MD and EPSR through an in-depth analysis of two widely used (and misused) drugs - Alprazolam and Cocaine. The results suggest that the Monte Carlo methodology inadequately samples inter-molecular interactions causing the solvation structures predicted by the two protocols to differ.

5.2 Introduction

5.2.1 Empirical Potential Structure Refinement

Neutron diffraction with its ability to resolve proton positions offers an experimental pathway to validate the computational modelling of ligand hydration [84–86, 158]. Additionally, with the scattering length of hydrogen ($-3.75fm$) different than that of deuterium ($6.67fm$) [159], isotopic substitution can be used to obtain different diffraction patterns for the same chemical system. Neutron diffraction patterns provide the static structure factor $F(Q)$ - a measure of the structure in reciprocal space [160]. $F(Q)$ can be written as:

$$F(Q) = \sum_{\alpha,\beta \geq \alpha} (2 - \delta_{\alpha\beta}) \cdot c_\alpha c_\beta b_\alpha b_\beta \left[S_{\alpha\beta} - 1\right] \tag{5.1}$$

where c_i and b_i are the relative concentration and scattering length of atom i respectively and $\delta_{\alpha\beta}$ is the Kronecker delta function. Q is the scattering vector defined by $Q = 4\pi/\lambda \cdot \sin(2\theta/2)$ where λ is the neutron wavelength and 2θ is the scattering angle. $S_{\alpha\beta}(Q)$ is the partial structure factor and is related to the Radial Distribution Function (RDF) $g_{\alpha\beta}(r)$ by the the Fourier Transform:

$$S_{\alpha\beta}(Q) = 1 + 4\pi\rho \int r^2 \cdot [g_{\alpha\beta}(r) - 1] \cdot \frac{\sin(Qr)}{Qr} dr \tag{5.2}$$

where ρ is the atomic number density of the sample. Thus, a relatively high density of solute in the solution is required to adequately sample its $S(Q)$.

Empirical Potential Structure Refinement (EPSR) [161] is a reverse Monte Carlo technique explicitly developed for the determination of local interactions from diffraction data ($F(Q)$). EPSR uses a box of molecules at the same concentration, density and temperature as the diffraction samples. A direct comparison to the

MD simulation studies at infinite dilution is thus not possible and a concentrated simulation box must be considered.

Figure 5.1 shows the pipeline of the EPSR simulation methodology. Starting '*seed*' potentials given to each unique atom consist of the parameters of the Lennard-Jones potential (equilibrium length σ and well-depth ϵ) and atomic charges (q_i). These parameters are used to carry out a simulation step before calculating the RDF $g_{\alpha\beta}(r)$. Structure factors from simulation ($F_{comp}(Q)$) are calculated through a Fourier transform of $g_{\alpha\beta}(r)$. Empirical potential, a function of the difference between experimental and computational structure factors is then calculated and applied to the simulation box. The steps are then repeated iteratively to fit the atomic conformation of the system to the experimental diffraction data.

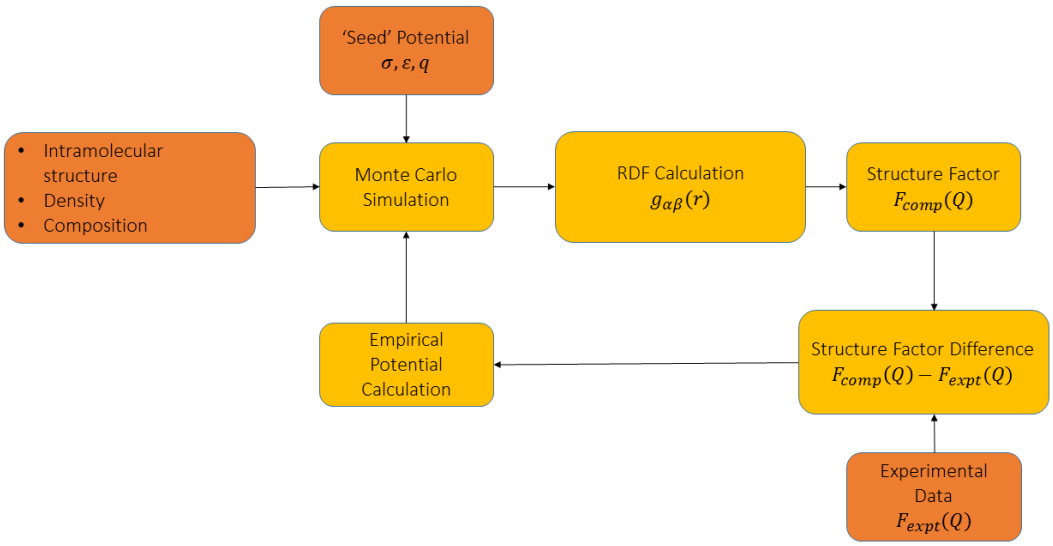


Figure 5.1 Schematic illustration of the EPSR simulation steps.

5.2.2 Molecules used for comparison

5.2.2.1 Alprazolam

Benzodiazepines are psychoactive small molecules that together are the most prescribed medications globally [162]. As the name suggets, the core chemical

structure of benzodiazepines is the fusion of benzene and diazepine rings. They generally have sedative, amnesic, and muscle relaxing properties and are used in the treatment of a range of disorders including anxiety, insomnia and epilepsy [163, 164]. The pharmacological effects of benzodiazepines are believed to result from their binding and modulating of the $GABA_A$ receptor [165, 166]. Additionally, benzodiazepines have also been discovered to bind to the BET (Bromodomain and Extra-terminal) [167] and Translocator (TSPO) family of proteins [168].

Alprazolam (Xanax®) is the most used (and misused) member of the benzodiazepine family of drugs [169]. It is also postulated to be the most toxic drug in the family [170] and has a long residence time [171]. In addition to the benzodiazepine targets GABA and BET, spectroscopic studies have predicted its binding to a range of important biomolecules including haemoglobin [172], albumin [173, 174] and even DNA [175]. Figure 5.2 shows the structure of Alprazolam (8-chloro-1-methyl-6-phenyl-4H-s-triazolo(4,3-a)(1,4)benzodiazepine) along with the naming terminology of salient atoms used in this work. Its structure consists of four distinct benzene, chlorobenzene, diazepine and triazole ring structures. With a predominantly hydrophobic surface, alprazolam is barely soluble in water ($\sim$ 0.13 mM) and is even predicted to interact with other hydrophobic drugs like fluoxetine [174].

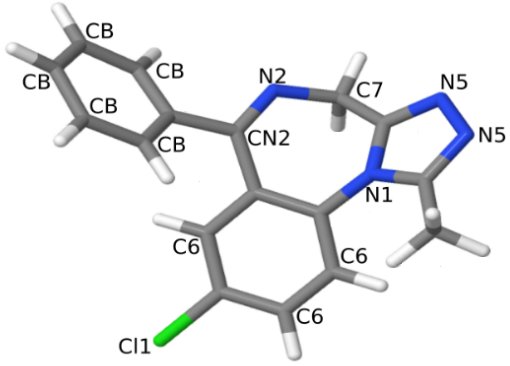


Figure 5.2 The molecular structure of alprazolam with the labelling of the salient atomic sites.

Despite the extensive usage of alprazolam and its diverse binding capability, limited structural information exists on its conformation and hydration both in bound and free state. A search of the Protein Data Bank [59] reveals only one structure of alprazolam bound to the first human BRD4 (PDB ID: **3U5J**) [129]. Regardless of its hydrophobicity, alprazolam binds in a hydrophilic pocket of BRD4 with two distinct bridging waters. Figure 5.3 shows the alprazolam binding site in the bromodomain and highlights the location of bridging waters. The water molecules play an important role in the binding of the drug by mediating its interactions with Tyr97 and Asn140 [13].

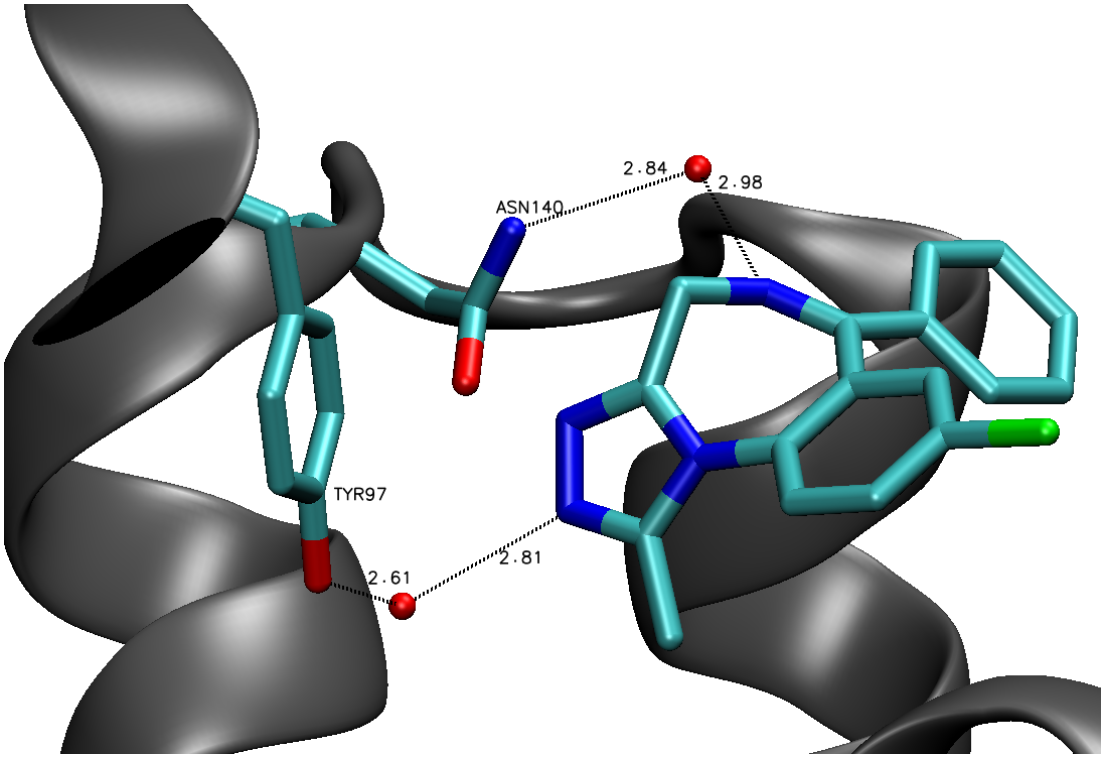


Figure 5.3 The binding configuration of alprazolam in the first human bromodomain BRD4 (PDB ID: **3U5J**) [129].

The importance of water molecules in its binding and the limited structural information available together make alprazolam an ideal candidate for in-depth solvation analysis.

5.2.2.2 Cocaine and its Stereo-isomers

Cocaine is an amphiphilic tropane alkaloid [138, 139] extracted from the leaves of the *Erythroxylum Coca* plant [176] and is the second largest drug of abuse worldwide [140]. It can cause potent cardiovascular and central nervous system stimulation [177] and the effects of use include euphoria, agitation, delirium, increased heart-rate and loss of thermoregulation [141, 142]. With its ability to bind and block sodium channels [178, 179], cocaine found some use as a topical anaesthetic [180, 181]. However, it is believed to function primarily by inhibiting dopamine reuptake after binding to the dopamine transport protein (DAT) and thus causing a build-up of dopamine in the synaptic cleft [182, 183]. Gabe et al. [184] first elucidated the structure of cocaine in 1962 after crystallization of its hydrochloride salt. Cocaine ($C_{17}H_{21}NO_4$) has been identified as methyl (1R,2R,3S,5S)-3-(benzoyloxy)-8-methyl-8-azabicyclo[3.2.1]-octane-2-carboxylate. Figure 5.4 shows the structure of cocaine with the chiral carbons marked as '*'.

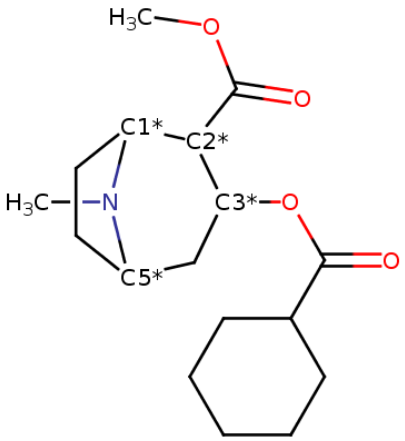


Figure 5.4 The molecular structure of cocaine with the chiral carbons labelled.

With steric stability necessitating carbons C1 and C5 have fixed and opposite configurations, cocaine has four diastereoisomers depending on the configurations of the carbons at atoms C2 and C3 [185]. Figure 5.5 shows the 3D structures of

the four stereoisomers (left pane) alongside the general atom-numbering scheme (right pane) followed in this work.

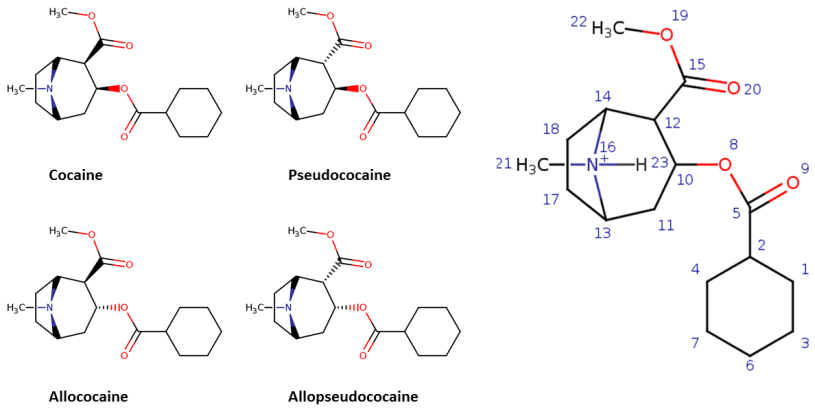


Figure 5.5 The four steroisomers of cocaine (left pane) and the atom numbering scheme (right pane).

A search of the Protein Data Bank [59] reveals six structures with cocaine bound to three different proteins - Drosophila dopamine transporter, ACh-Binding Protein and a catalytic antibody. Table 5.1 lists the crystallographic structures with cocaine bound as a ligand. Four of the six crystallographic structures have a water molecule subtended between the amine and methoxy-ester groups. Figure 5.6 shows the cocaine binding site in the GNC92H2 antibody (**1I7Z**) [144] and highlights the location of the bridging waters molecule.

Table 5.1 Cocaine bound crystallographic Structures in the Protein Database.

PDB ID	Resolution (Å)	Protein
4XPB	3.05	Dopamine Transporter
4XP4	2.80	Dopamine Transporter
2PGZ	1.76	ACh-Binding Protein
2AJV	1.50	Catalytic Antibody
1Q72	1.70	M82G2 Antibody
1I7Z	2.30	GNC92H2 Antibody

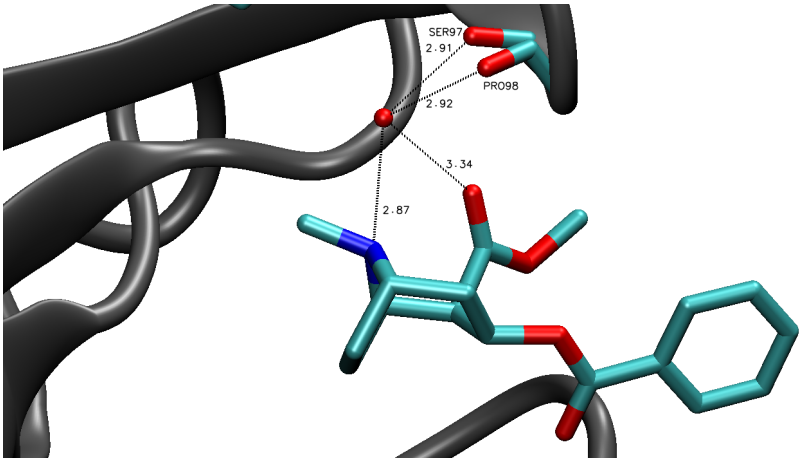


Figure 5.6 The binding configuration of cocaine in the GNC92H2 Antibody (PDB ID: **1I7Z**) [144].

The ubiquitous nature of the water molecule and cocaine's pharmacological relevance makes its hydration analysis important. While Diffraction and EPSR studies were only carried out on cocaine, MD simulations were also performed on its stereoisomers to help identify the structural basis for its behaviour in solution.

5.3 Methods

5.3.1 Neutron Diffraction and EPSR[1]

5.3.1.1 Alprazolam

Alprazolam (CAS 28981-97-7) was purchased from Sigma-Aldrich, verified by ^{1}H NMR and used without further purification. With its low solubility in water, the solvation of Alprazolam was analyzed in a 50-50 water-methanol solution. Six water-methanol solutions of different isotopomers were used to obtain different diffraction patterns. The isotopomers of methanol were initially dried over Mg that had been previously activated with I_2. The isotopomers were then refluxed at 330 K for 48 hours under vacuum before being cryogenically distilled onto

[1]All work in this section was performed by Andrew Johnston from the McLain Research Group

pre-dried 3 Å molecular sieves. The anhydrous methanol(s) were finally verified through ^{1}H NMR. Finally, the alprazolam-water-methanol solutions were prepared at a relative molar ratio of 1:125:125 (0.14 M) under an N_2 atmosphere by weight. To further promote miscibility, the solution was warmed to 323 K and kept at this temperature throughout the diffraction measurement. Table 5.2 lists the solvent compositions of the six samples and the name used to refer each sample in subsequent plots.

Table 5.2 Solvent compositions of the six alprazolam solution samples.

Sample Number	Sample Name	Methanol	Water
1	alprazolam/CH_3OH/H_2O	CH_3OH	H_2O
2	alprazolam/CHD_3OH/H_2O	$CH_{1.5}D_{1.5}OH$	H_2O
3	alprazolam/CH_3OD/D_2O	CH_3OD	D_2O
4	alprazolam/CD_3OD/D_2O	CD_3OD	D_2O
5	alprazolam/CHD_3OD/D_2O	$CH_{1.5}D_{1.5}OD$	D_2O
6	alprazolam/CD_3OHD/HDO	$CD_3OH_{0.5}D_{0.5}$	HDO

Diffraction data was collected for between 6 to 8.5 hours per sample. The data was then corrected for absorption, multiple scattering and inelasticity effects and subsequently converted to $F(Q)$ using the GUDRUN program [186].

EPSR simulation for fitting to the $F(Q)$ data was performed on a modelling box of 25 Alprazolam, 3125 water and 3125 methanol molecules at 323 K. The 'seed' parameters of alprazolam and methanol were generated using the *tleap* and *Antechamber* [108] modules of Amber14 [109] before slight modification to account for the EPSR atom-naming conventions. The parameters for water were taken from the SPC/E water model [96].

5.3.1.2 Cocaine

Cocaine hydrochloride (CAS 53-21-4) was purchased from Sigma Aldrich, verified by ^{1}H NMR and used without further purification. Cocaine-HCl was deuterated by dissolution in 99.9% D_2O and the solvent was subsequently removed by vacuum. This process was repeated three times to ensure adequate deuteration. The success of this deuteration was verified by ^{1}H NMR. All isotopomeric solutions of cocaine-HCl in water for neutron diffraction experiments were prepared at a 1:65 ratio (0.8 M) by weight under an N_2 atmosphere. Table 5.3 lists the isotopic composition of cocaine and water in the six samples.

Table 5.3 Solvent compositions of the six cocaine solution samples.

Sample Number	Cocaine	Water
1	h-cocaine	H_2O
2	$h_{0.75}d_{0.25}$-cocaine	$H_{0.75}D_{0.25}O$
3	$h_{0.50}d_{0.50}$-cocaine	$H_{0.50}D_{0.50}O$
4	$h_{0.64}d_{0.36}$-cocaine	$H_{0.64}D_{0.36}O$
5	d-cocaine	D_2O
6	$h_{0.25}d_{0.75}$-cocaine	$H_{0.25}D_{0.75}O$

Unlike alprazolam, the solvation of cocaine did not require warming or the addition of methanol to the solution. Diffraction data was collected for between 8 to 9.5 hours per sample. The data was then corrected for absorption, multiple scattering and inelasticity effects and subsequently converted to $F(Q)$ using the GUDRUN program [186].

EPSR simulation for fitting to the $F(Q)$ data was performed on a modelling box of 25 protonated cocaine, 25 Cl^- and 1625 water molecules at 298 K. The 'seed' parameters of cocaine and Cl^- were generated using the *tleap* and *Antechamber* [108] modules of Amber14 [109] before slight modification to account for the EPSR

atom-naming conventions. The parameters for water were taken from the SPC/E water model [96].

5.3.2 Molecular Dynamics Simulation

5.3.2.1 Alprazolam

Two sets of Molecular Dynamics simulations were performed to validate the EPSR results and compare their performance in modelling solvation. The simulations were performed at the same molecular ratio as EPSR - 25 alprazolam, 3125 water and 3125 methanol molecules. The atomic partial charges of alprazolam and methanol were initially calculated according to the AM1-BCC method [102, 103] using *Antechamber*. Subsequently, the molecules were parameterized according to the GAFF [81] force-field using Amber14. Finally, the parameters were converted to Gromacs format using the *acpype* script [116]. Water was modelled using the SPC/E model [96] for the Amberff99SB-ILDN force-field [79].

First, the density of the two solvents were used to calculate the dimensions of a cubical box containing 3125 water and 3125 methanol molecules. A random orientation of 25 alprazolam molecules was placed in this box and the solvent molecules were subsequently added. Energy Minimization (EM) corrected atomic overlaps up to the point where the maximum force on any atoms was below 100 $kJmol^{-1}nm^{-1}$. EM was followed by solute restrained equilibration runs of 1 ns each in the NVT and NPT ensembles. The Berendsen thermostat and barostat [95] were used to maintain temperature and pressure in the equilibration runs. One solution box was simulated at a temperature similar to EPSR (323 K). The second simulation was carried out at a physiological temperature of 300 K while the pressure was 1 bar in both runs.

The production runs in both simulation boxes were carried out in the NPT ensemble for 40 ns with a timestep of 2 fs. Temperature was maintained using the

V-rescale thermostat while pressure was controlled using the Parinello-Rahman barostat [118]. Long range interactions were calculated using the particle mesh Ewald (PME) algorithm [93, 94]. Atomic co-ordinates were saved every 0.8 ps, resulting in 50000 simulation frames.

5.3.2.2 Cocaine and its stereo-isomers

The structure of cocaine and its three diastereoisomers were downloaded from the PubChem database [187] in *sdf* format and converted to *mol2* format using *OpenBabel* [188]. Table 5.4 lists the PubChem Compound ID (CID) of all four structures.

Table 5.4 The four cocaine diastereoisomers and their PubChem CID.

Stereo-isomer	PubChem CID
Cocaine	446220
Pseudococaine	9861246
Allococaine	15558682
Allopseudococaine	638023

The partial atomic charges of all isomers were generated according to the AM1-BCC method using *Antechamber*. Subsequently, the molecules were parameterized according to the GAFF force-field using Amber14. Finally, the parameters were converted to Gromacs format using the *acpype* script. The TIP3P [82] parameters for the Amberff99SB-ILDN [79] force-field were used to model the water molecules. The partial atomic charges assigned by *Antechamber* to cocaine and its diastereoisomers were then verified to ensure no major discrepancies (see Table 5.5). This uniformity of partial charges ensure that any differences in hydration structure and inter-molecular interactions are only due to atomic spatial orientation.

Table 5.5 Atomic charge dispersal within the cocaine diastereoisomers.

Atom No.	Element	Cocaine	Pseudococaine	Allococaine	Allopseudococaine
1	C	-0.0645	-0.0650	-0.067	-0.0680
2	C	-0.1536	-0.1546	-0.1546	-0.1586
3	C	-0.1415	-0.1410	-0.1390	-0.1400
4	C	-0.0645	-0.0650	-0.067	-0.0680
5	C	0.6697	0.6677	0.6607	0.6617
6	C	-0.0740	-0.0750	-0.0750	-0.0760
7	C	-0.1415	-0.1410	-0.1390	-0.1400
8	O	-0.4489	-0.4519	-0.4659	-0.4639
9	O	-0.5370	-0.5360	-0.5250	-0.5270
10	C	0.1611	0.1531	0.1571	0.1581
11	C	-0.1414	-0.1434	-0.1294	-0.1344
12	C	-0.1447	-0.1457	-0.1467	-0.1417
13	C	0.1445	0.1455	0.1415	0.1435
14	C	0.1475	0.1485	0.1495	0.1525
15	C	0.6471	0.6341	0.6271	0.6471
16	N	-0.7114	-0.7024	-0.7084	-0.7044
17	C	-0.1054	-0.1024	-0.0984	-0.0974
18	C	-0.1024	-0.1124	-0.0964	-0.0974
19	O	-0.3939	-0.4279	-0.3979	-0.4249
20	O	-0.5770	-0.5050	-0.5710	-0.5160
21	C	0.0911	0.0881	0.0921	0.0891
22	C	0.1197	0.1217	0.1207	0.1237
23	H^+	0.4898	0.4548	0.4838	0.4478

The hydration and inter-molecular interactions of cocaine and its diastereoisomers were investigated at a range of concentrations. First, a random orientation of 64 solute molecules was placed in a simulation box. An equal number of Cl^- ions

were added to neutralize the system followed by solvation in water. The ratio of number of molecules of stereoisomers to water in the solution varied from 1:50 to 1:1000. In addition, the hydration shell at infinite dilution (excluding inter-solute interaction) was calculated by placing the cocaine diastereoisomers at the centre of a box and 1.2 nm from the edges. This box was then solvated in TIP3P water and neutralized by Cl^- ions.

Atomic overlaps were corrected by EM up to the point where the maximum force on any atom was below 100 kJ mol^{-1} nm^{-1}. EM was followed by a solute restrained equilibration runs in the NVT and NPT ensembles of 1 ns each. The temperature and pressure were maintained at 300 K and 1 bar respectively using the Berendsen algorithm [95]. The production runs were carried out in the NPT ensemble for 30 ns with a timestep of 2 fs. Temperature was maintained using the V-rescale thermostat while pressure was controlled using the Parinello-Rahman barostat [118]. Long range interactions were calculated using the particle mesh Ewald (PME) algorithm. Atomic co-ordinates were saved every 0.6 ps, resulting in 50000 simulation frames. Table 5.6 lists the details of the MD simulations - the solute:water ratios and the corresponding average box volume for each of the runs.

Table 5.6 Cocaine stereoisomer simulation particulars.

Stereo-isomer	Solute Number	Solvent Number	Box Volume (nm^3)
Cocaine	64	3200,6400,32000,64000	130,220,997,1967
Pseudococaine	64	3200,6400,32000,64000	131, 221, 1000, 1972
Allococaine	64	3200,6400,32000,64000	132, 222, 1000, 1971
Allopseudococaine	64	3200,6400,32000,64000	129, 221, 998, 1971
Cocaine	25	1625	53.06
Cocaine	1	1400	44.89
Pseudococaine	1	1400	44.66
Allococaine	1	1400	42.39
Allopseudococaine	1	1400	41.89

5.3.3 Spatial Density Analysis

The behaviour of liquids in simulation is traditionally studied through Radial Distribution Functions (RDF - $g(r)$) [67]. However, $g(r)$ plots are radially averaged and thus do not help identify site-specific interactions or orientational behaviour of molecules [88]. Hence, a visualization of the three-dimensional atomic arrangement is imperative especially in concentrated solutions like those in this work.

The program ANGULA [189] was used to extract the three-dimensional arrangements of molecules relative to one another from the EPSR simulation box. In ANGULA, orthonormal coordinate systems are assigned to different fragments of the solute and solvent molecules. The angles between the assigned orthonormal axes are then used to find the directionality of the interactions. Finally, a 3-dimensional histogram of the solvent molecule distribution around solute fragments is plotted to find the magnitude of interactions.

Python MDAnalysis [119] scripts utilizing the 3D rotation matrix algorithm [190] were used to identify the three-dimensional arrangement of molecules from MD simulations. Fragments from all solute molecules in every trajectory frame were iteratively mapped onto a fictional solute molecule placed at the origin. The rotation matrix of each mapping was then iteratively applied to transform the co-ordinates of all solvent molecules neighbouring the solute fragment. Finally, the 'transformed' co-ordinates were histogrammed to provide a 3-dimensional distribution of interactions.

The histogrammed 3D distributions were plotted as Spatial Density Maps (SDM) using the *scipy*, *matplotlib* [191] and *mayavi* [192] libraries.

5.4 Results

5.4.1 Alprazolam

5.4.1.1 Fitting to Experimental Diffraction Data

Figure 5.7 shows the measured $F(Q)$ diffraction data along with a comparison to the $F(Q)$ functions derived from the EPSR (left-pane) and MD (right-pane) simulation boxes. The measured data, simulation fits and the difference between the two are shown as coloured, black and grey lines respectively. Figure 5.8 plots a similar comparison of the data albeit in real space obtained through a Fourier transform of the data in Figure 5.7.

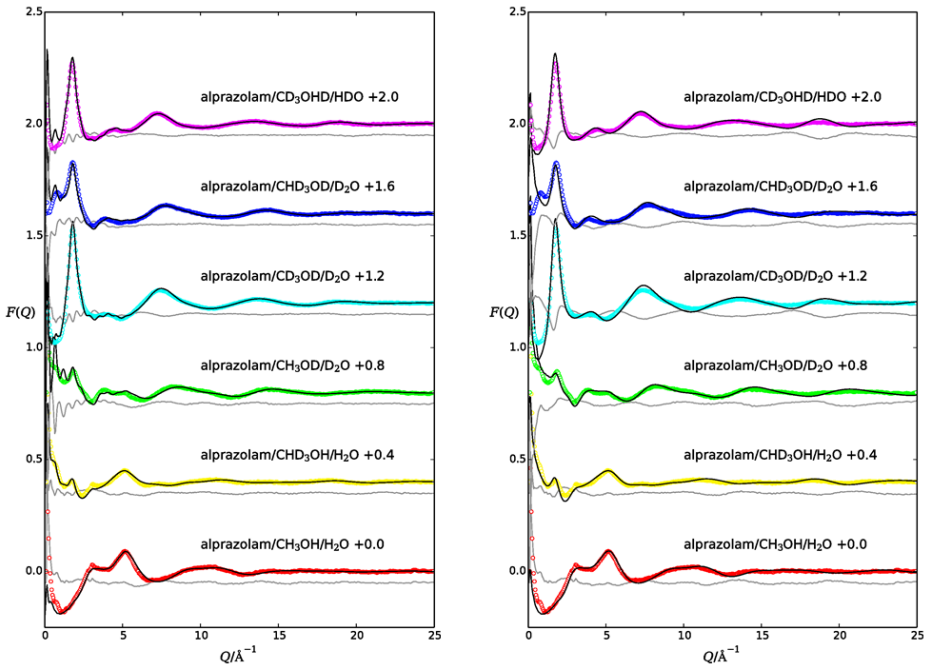


Figure 5.7 Plot showing the comparison of the measured $F(Q)$ (coloured lines) with the EPSR (black lines - left pane) and MD (black lines - right pane) fits to data. The differences between the data are shown as gray lines.

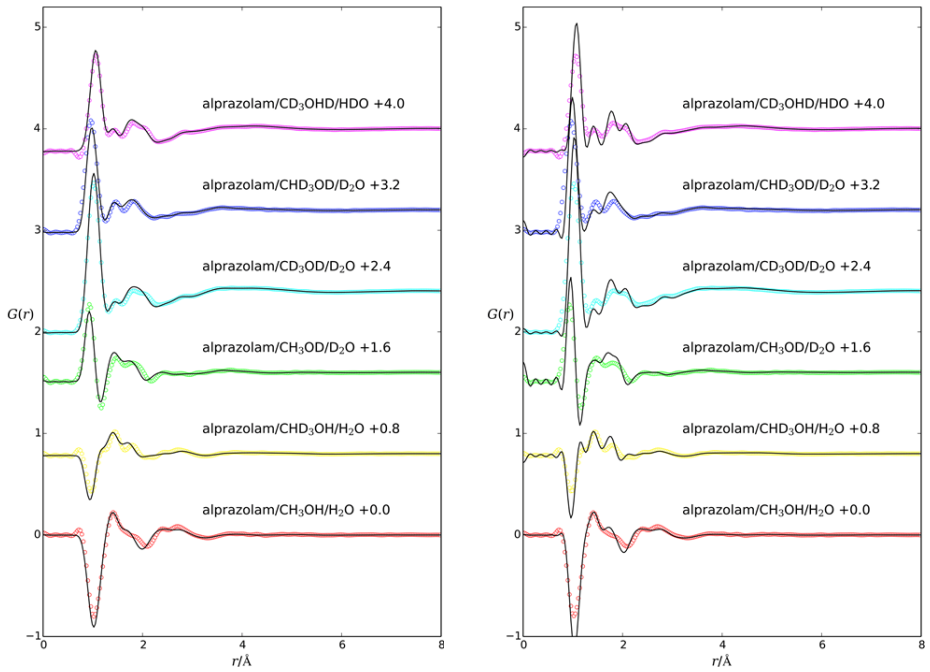


Figure 5.8 Plot showing the $F(Q)$ comparison for alprazolam in real space with EPSR plotted in the left pane and MD in the right pane. The experimental data is plotted as coloured lines and the data calculated from simulation are plotted as black lines.

The EPSR and MD fits to the measured $F(Q)$ data are both generally good with only small differences occurring at small Q values. Previous work have attributed this difference to the inherent difficulty in correcting for inelastic scattering in this region of data [193]. The maximum deviation of MD data from the experimental results occurs in the fully deuterated sample 4 (CD_3OD/D_2O). With hydrogen and the heavy atoms having opposite scattering directions [159], maximum deviation might be observed in this sample probably due to the least cancelling out of errors [194, 195].

5.4.1.2 Methanol interactions with alprazolam

The interactions of methanol with the each of the four ring structures of alprazolam predicted though both MD and EPSR is plotted in Figure 5.9. Figures 5.9 a-d show the atom-averaged RDF of the methanol O atoms around each of the four rings of alprazolam. The locations of the nearest neighbour methanol oxygen around each of the ring centres from EPSR (Figures 5.9 e-h) and MD (Figures 5.9

i-l) are plotted as SDMs. The agreement between the methanol solvation predicted by the two methods is extremely good. Additionally, the effects of temperature seem negligible with the MD $g(r)$ plots at 323 K and 300 K identical.

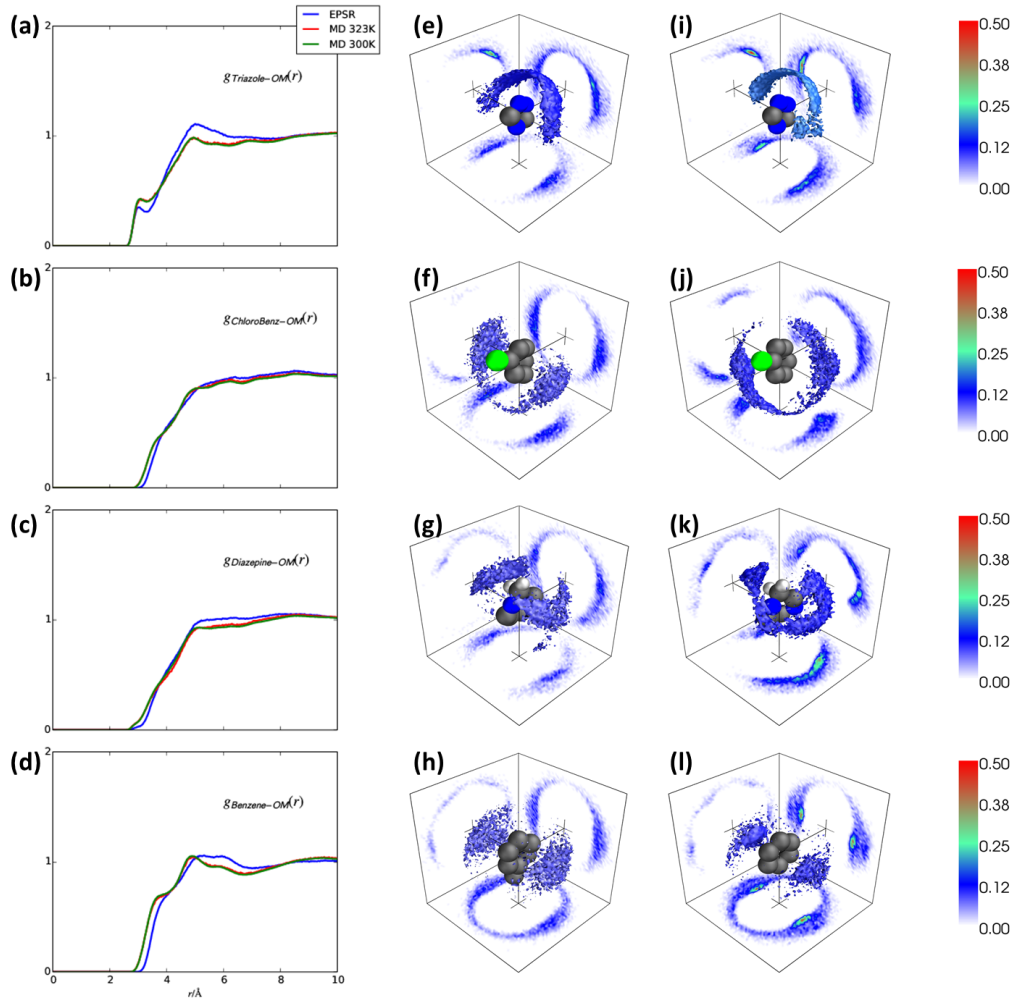


Figure 5.9 (a-d) RDFs of the oxygen atom of methanol around each of the four ring structures of alprazolam from EPSR, 300 K and 323K MD. **(e-h)** SDMs of the nearest methanol oxygen around the corresponding ring structures from EPSR. **(i-l)** SDMs of the nearest methanol oxygen around the corresponding ring structures from MD at 323 K. The isocontours enclose the densest 30% of sites and the scale bar shows the number density in $Å^{-3}$.

The $g(r)$ of methanol around the triazole ring describes two minor peaks at $\sim$ 3.0 Å and $\sim$ 5.5 Å (Figure 5.9a). The first peak at $\sim$ 3.0 Å is probably caused by the methanol OH group hydrogen bonding with the N5 atom as it is close to the O-H–N hydrogen bond length [196]. MD slightly over-estimates interactions compared with EPSR and its SDM (Figure 5.9i) correspondingly shows a greater

density when compared to EPSR (Figure 5.9e). The second peak at at $\sim$ 5.5 Å is close to the observed distance for the second solvation shell [197]. With MD predicting a greater number of methanol molecules in the first shell, its prediction for the methanol density in the second shell is correspondingly lower.

The $g(r)$s of methanol around the chloro-benzene and diazepine rings show no discernible peaks in both MD and EPSR (Figure 5.9 b,c). This suggests that methanol molecules have no preferred orientations around these ring structures. In the SDMs correspondingly, the contours describing the positions of methanol are dispersed with limited regions of high density (Figure 5.9 f,g,j,k).

The $g(r)$ of methanol around the benzene ring displays a small peak at $\sim$ 5 Å (Figure 5.9d). Like in the case of triazole, this peak is slightly more pronounced in MD than in EPSR. The SDM contours describe a preference for methanol to reside in a position normal to the plane of the aromatic ring. Additionally, the slightly more prominent peak in MD simulations is evident in the marginally greater density of its SDM (Figure 5.9l) compared to EPSR (Figure 5.9h).

5.4.1.3 Water interactions with alprazolam

Figure 5.10 describes the interactions of water with each of the four ring systems of alprazolam. Figures 5.10 a-d show the atom-averaged RDF of the water O atoms around each of the four rings of alprazolam. The locations of the nearest neighbour water oxygen around each of the ring centres from EPSR (Figures 5.10 e-h) and MD (Figures 5.10 i-l) are plotted as SDMs. While qualitatively similar, the EPSR and MD predictions diverge to a greater extent quantitatively when compared to the previous methanol analysis (Figure 5.9). EPSR consistently predicts greater hydration across all ring structures of alprazolam.

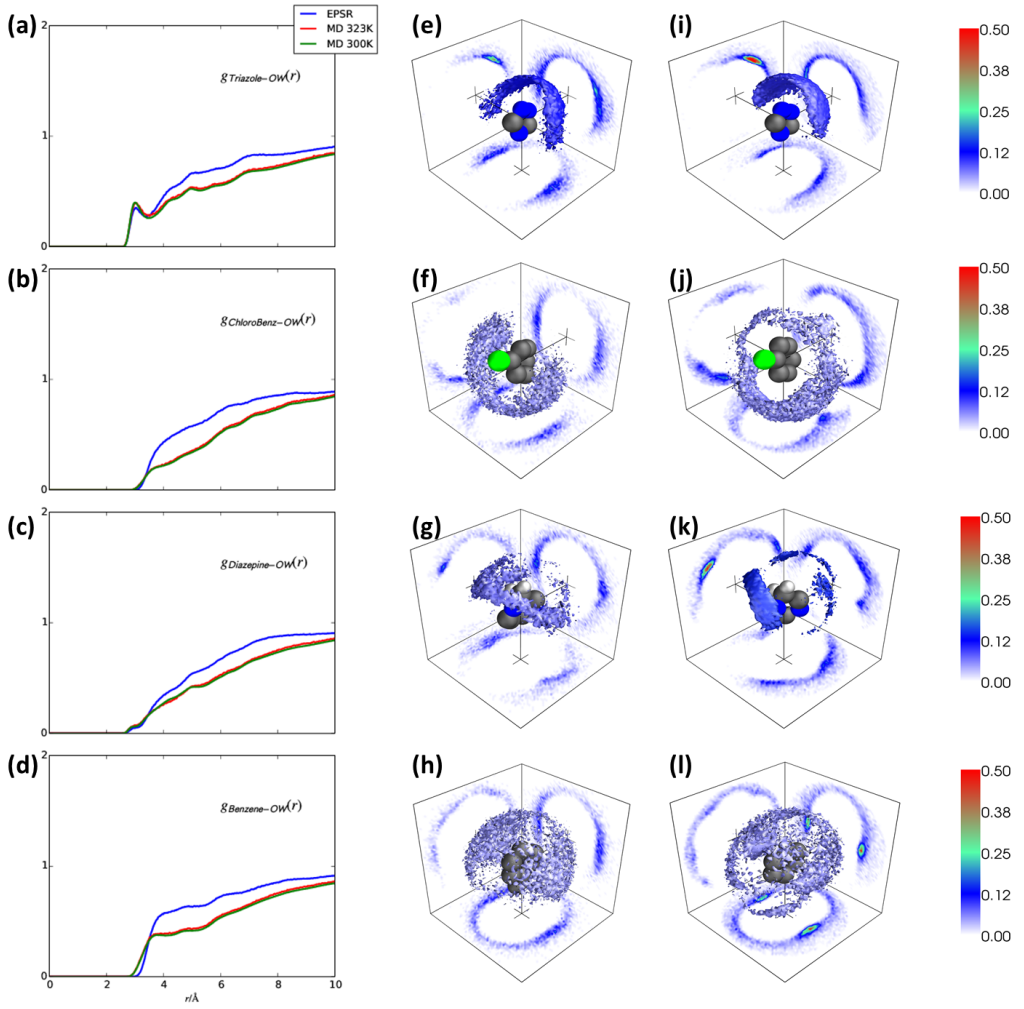


Figure 5.10 (a-d) RDFs of the oxygen atom of water around each of the four ring structures of alprazolam from EPSR, 300 K and 323K MD. **(e-h)** SDMs of the nearest water oxygen around the corresponding ring structures from EPSR. **(i-l)** SDMs of the nearest water oxygen around the corresponding ring structures from MD at 323 K. The isocontours enclose the densest 30% of sites and the scale bar shows the number density in $Å^{-3}$.

Similar to the methanol solvation, the hydration of triazole shows a peak at 3.0 Å (Figure 5.10 a). This N–H-O hydrogen bond is also reflected in the SDM map which displays a sharp density profile around the two N5 atoms (Figure 5.10 e,i). However, in contrast to methanol, both EPSR and MD display no second peak and thus minimal water presence is observed in the second solvation shell.

The $g(r)$s of water around both aromatic rings show no peak (Figure 5.10 b,d). Consistently, the SDM contours of water around them are diffuse with few preferential orientations (Figure 5.10 f,j,h,l). This diffuse density with little

localization might be due to the triazole ring structure offering greater electrostatic H-bonding opportunities and thus reducing water-benzene interactions.

The hydration structure of the diazepine ring presents a contrast between the RDF and SDM predictions. Despite the $g(r)$ plots displaying no major peaks (Figure 5.10d), the SDM contours indicate directional interactions with the N2 atom (Figure 5.10 h,l). Additionally, the interaction is comparatively more directional in MD. The dissimilarity between the SDM and RDF might probably be due to the N2-Ow interactions being averaged out over the seven atoms in the diazepine ring.

5.4.1.4 Atomistic Preferential Solvation

Figure 5.11 further examines the interaction of methanol and water with alprazolam by plotting the RDFs for specific atoms. In general, the MD and EPSR results again agree with each other and also support the SDM contours in Figures 5.9 and 5.10. Specifically, the N2-Ow graph explains the localized water distribution seen around the diazepine ring (Figure 5.10 h,l) with the plot exhibiting a distinct peak at $\sim$3 Å (h-bond distance) [196]. The greater localization of water around the diazepine ring seen in MD (Figure 5.10 h,l) is also evident with the MD N2-Ow plot exhibiting a larger and sharper peak compared to EPSR.

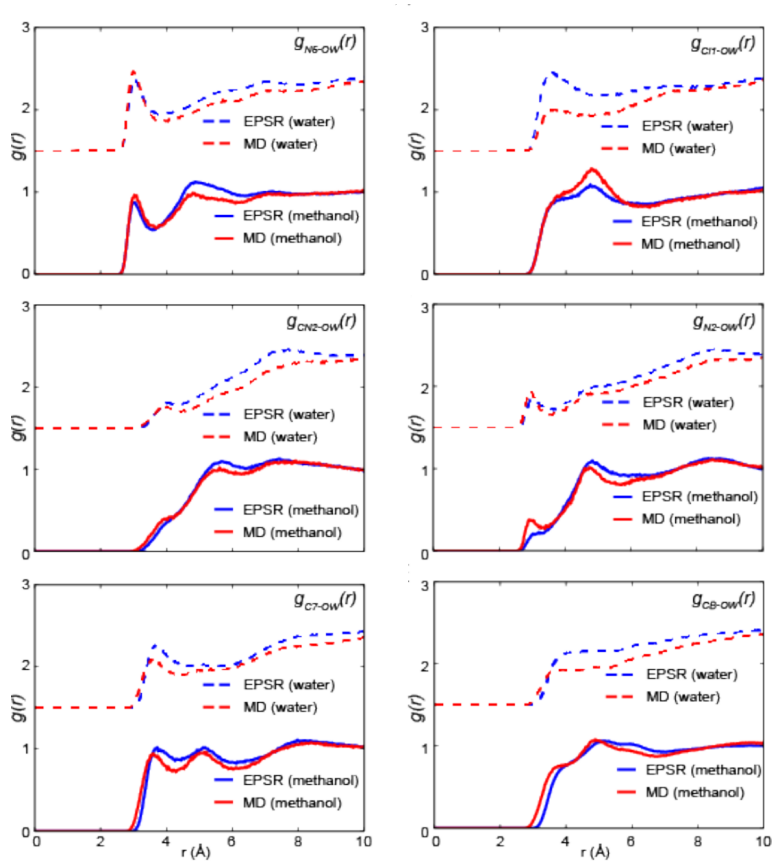


Figure 5.11 Radial Distribution Functions from MD and EPSR of Methanol and Water Oxygen around salient atomic sites of alprazolam. The RDFs for water have been displaced by 1.5 units for clarity.

To further quantify the hydrophobic/hydrophilic balance of Alprazolam, the preference ratio is calculated. The preference ratio P_{met}^{wat} calculates the preferential interactions of an alprazolam atom for water over methanol. It is defined as

$$P_{met}^{wat} = \frac{n_x^{Ow}(r)}{n_x^{Om}(r)} \tag{5.3}$$

where n_x^{Ow} and n_x^{Om} are the co-ordination numbers of the water and methanol Oxygens respectively around atom x at radius r. A ratio greater than 1 indicates a preference of water while a value lesser than 1 indicates the opposite. Preference ratio calculations have previously been used in EPSR and MD studies of the competitive interactions of urea with polypeptides [198, 199]. To encompass the first solvation shell, the co-ordination numbers of water and methanol are calculated at the first minimum of the respective atom's RDF plots in Figure 5.11.

Table 5.7 lists the n_x^{Ow} and n_x^{Om} co-ordination numbers of the alprazolam atoms, the r values at which they were calculated and their corresponding preference ratios.

Table 5.7 Co-ordination numbers and preference ratios of alprazolam atoms from both MD and EPSR. The r values for the calculation of the co-ordination numbers are taken as the first minima of the RDF graphs in Figure 5.11 to completely encompass the first hydration shell.

Atom	r(Å)	n^{Ow} MD–EPSR	n^{Om} MD–EPSR	P_{met}^{wat} MD–EPSR
N5	3.84	0.86 – 0.88	1.01 – 0.95	0.851 – 0.926
C7	4.38	1.04 – 1.28	1.74 – 1.81	0.598 – 0.707
N2	3.44	0.25 – 0.24	0.29 – 0.16	0.862 – 1.500
CN2	4.42	0.43 – 0.48	0.73 – 0.62	0.589 – 0.774
CL1	4.68	1.36 – 2.41	2.90 – 2.68	0.469 – 0.899
CB	4.24	0.74 – 0.90	1.30 – 1.06	0.569 – 0.849

The coordination numbers for MD and EPSR are broadly similar with MD predicting slightly greater hydration of polar atoms than EPSR. With the limited solubility of alprazolam in water, most atoms have a preference ratio < 1, indicating a preference for methanol over water. However, the solvent exposed N5 and N2 atoms interact the most with water and as such have the largest preference ratios. Additionally, the directionality of water around the N2 atom seen in Figure 5.10 is also corroborated with the adjacent C7 and CN2 atoms having markedly lower preference ratios.

5.4.1.5 Comparing the bound and solution states

5.4.1.5.1 Conformation

The structure of alprazolam imposes a degree of rigidity on it. With fused chlorobenzene, diazepine and triazole rings, the benzene ring remains the most mobile in solution. The conformational analysis of the EPSR simulation carried

out using ANGULA further supports this. An overlay of the different Alprazolam conformations observed over the course of the EPSR simulation is shown in Figure 5.12 a. Unlike MD, EPSR has limited bond constraints on atomic motion. Despite this, the figure highlights the limited movement of the triazole, diazepine and chlorobenzene rings. In contrast, the benzene ring is relatively mobile and is able to sample an array of conformations.

In the crystallographic structure of alprazolam bound to the BRD4 bromodomain protein [129], the benzene ring is only slightly tilted away from the plane of the molecule. In modelling studies however, Dangkoob et al. [174] predicted the benzene ring to adopt a conformation completely normal to the plane of the rest of the molecule. Figure 5.12 b plots the distribution of the dihedral angles of the rotatable CN2-CB1 bond from both MD and EPSR. With only two major peaks at -30° and 150°, the benzene ring adopts a slightly tilted configuration similar to its bound crystallographic structure where the angle is $\sim 28°$. Unlike MD, there is no complete rotation of the benzene ring in EPSR. However, the benzene ring again adopts a conformation slightly tilted to the molecular plane with peaks at 30° and $-30°$.

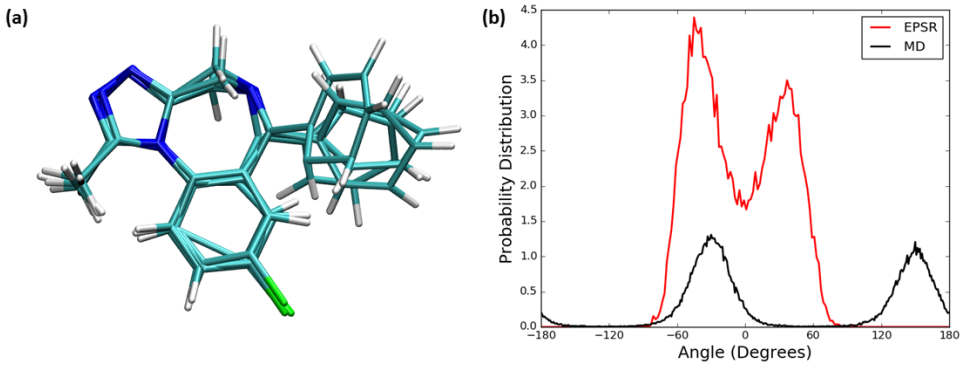


Figure 5.12 (a) Overlay of four different alprazolam molecules obtained from WMA of EPSR to demonstrate the conformations adopted by the aromatic ring. **(b)** Distribution of dihedral angles subtended by the rotation of the CN2-CB1 bond in the MD simulations at 323 K.

5.4.1.5.2 Solvation

Figure 5.13 plots the results of the MD Whole Molecule Analysis (WMA) of

the methanol and water shells of alprazolam. Molecules from 5000 simulation frames are overlaid onto the crystallographic alprazolam conformation and solvent molecules within a distance range of 0-4 Å were considered. The WMA thus represents the most probable solvent location around all atoms of alprazolam. The figure plots the contour occupied by the top 35% of solvent molecules - water in green and methanol in grey. Additionally, the crystallographic bridging waters (PDB ID: 3U5J) [129] are added as red spheres. Despite the marginal errors caused by the overlaying different conformations, the WMA hydration shell closely matches the atomic preferential solvation ratios of Table 5.7. The water contour shell surrounds the N5 and N2 atoms while methanol itneracts with the remaining hydrophobic motifs. In further support of the previous analysis described in Chapter 3, the preferential water orientations match the locations of bridging waters from the alprazolam-BRD4(1) complex.

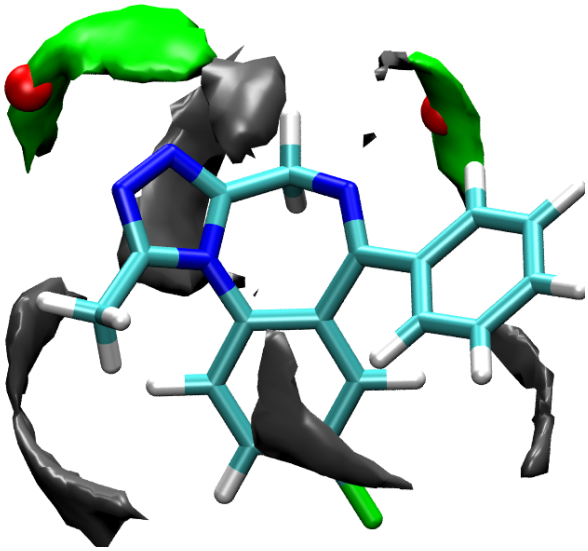


Figure 5.13 Whole Molecule Analysis of MD simulations at 323 K with the green and grey solvent clouds showing the probable locations of the top 35 % of water and methanol molecules respectively. The crystallographic bridging water molecule sites (PDB ID: 3U5J) are shown as red spheres.

5.4.2 Cocaine

5.4.2.1 Fitting to Experimental Diffraction Data

Figure 5.14 plots show the measured $F(Q)$ diffraction data along with a comparison to the functions derived from the EPSR (left-pane) and MD (right-pane) simulation boxes at the same molecular ratio. The measured data, simulation fits and the difference between the two are shown as coloured, black and grey lines respectively. Figure 5.15 plots a similar comparison in real-space obtained through a Fourier transform of the data in Figure 5.14.

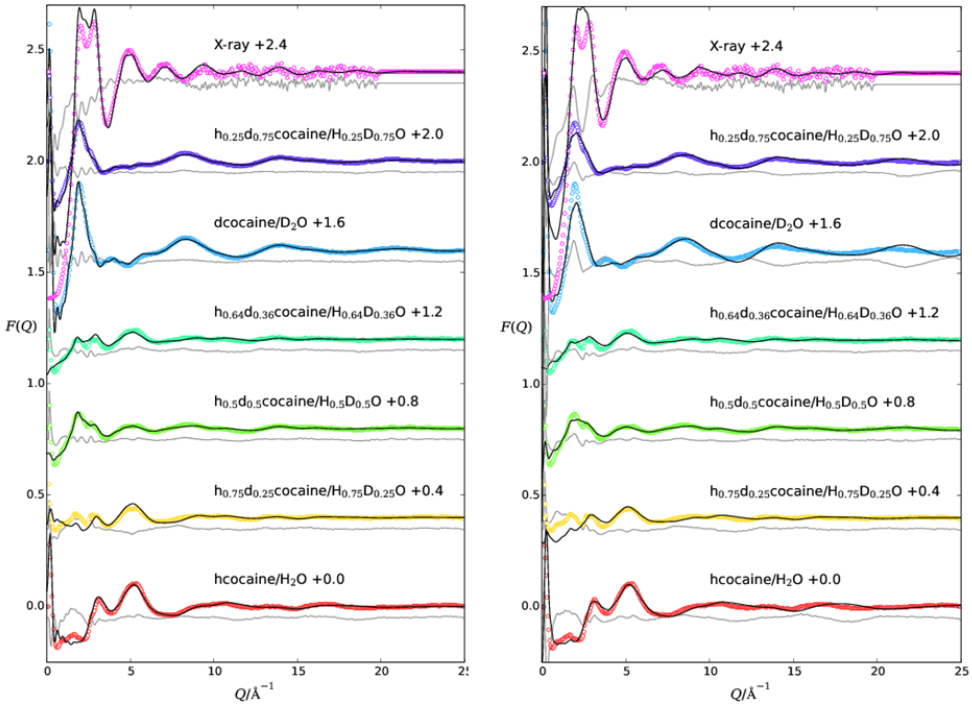


Figure 5.14 Plot showing the comparison of the Measured $F(Q)$ (coloured lines) with the EPSR (black lines - left pane) and MD (black lines - right pane) fits to data. The differences between the data are shown as gray lines.

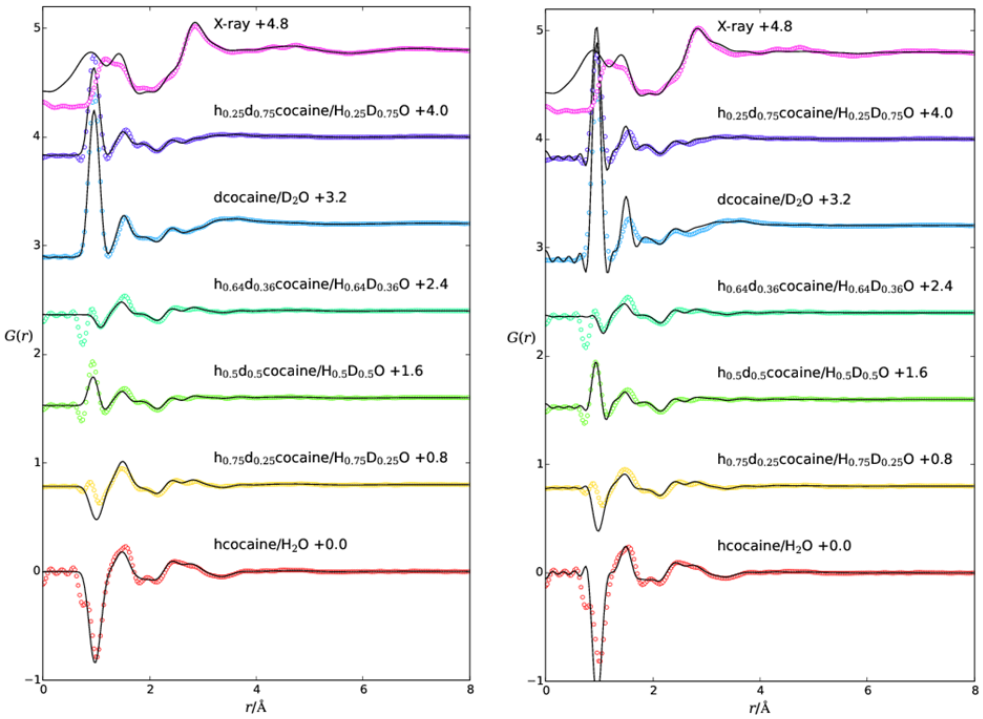


Figure 5.15 Plot showing the $F(Q)$ comparison for cocaine in real space with EPSR plotted in the left pane and MD in the right pane.

In the case of cocaine, EPSR again provides a better fit to the experimental data when compared to MD. The deviations of MD from the measured data is again maximum in the samples with uniform deuteration. However unlike alprazolam, substantial deviations of MD from the measured data are evident even in mixed isotopomeric mixtures with the cancelling out effect. This suggests that the atomic distributions in the two methodologies are significantly different.

5.4.2.2 Behaviour of Cocaine from EPSR[2]

Figure 5.16 succinctly summarizes the behaviour of cocaine in aqueous solutions from EPSR. Different cocaine conformations observed over the course of the simulation are overlaid to describe the relative mobility of the groups. Additionally, the most probable location of water molecules around cocaine is calculated using ANGULA Whole Molecule Analysis (WMA) and plotted as an iso-surface contour.

[2]Published as *On the atomic structure of cocaine in solution, Physical Chemistry Chemical Physics, 2016, 18, 991-999* and briefly discussed here for background

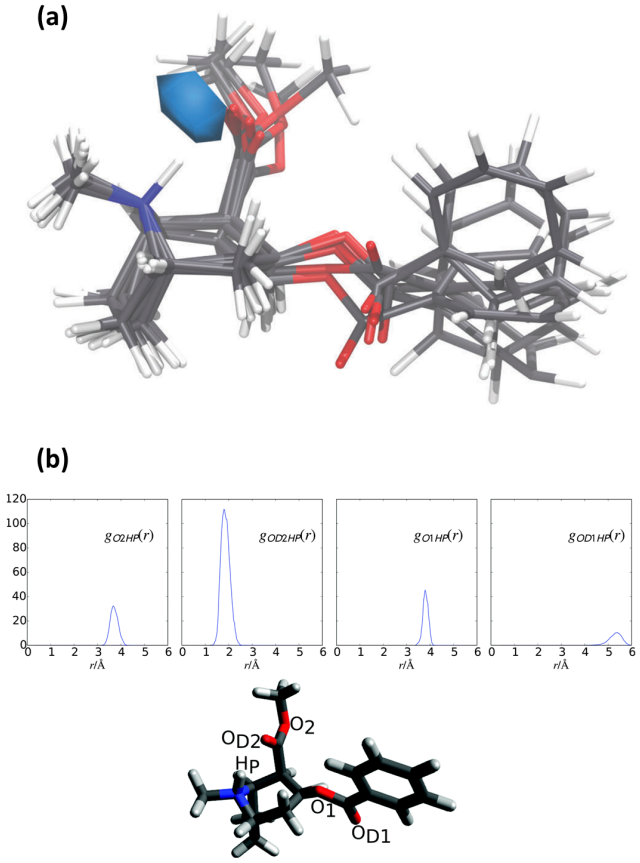


Figure 5.16 (a) Overlay of the Cocaine conformations in the EPSR simulation and a contour describing the most probable water location around the whole Cocaine molecule. **(b)** Intra-molecular RDF functions of cocaine atoms. Additionally, the atom naming terminology in the RDF plots is elaborated. Image courtesy Johnston et al. [89]

Figure 5.16a shows that the tropane ring remains rigid throughout the simulation. However, despite the limited constraints placed on atomic structure in EPSR, the methoxy-ester group also remains relatively rigid. The benzoxy ester remains the only mobile motif in solution and samples the widest array of conformations. Additionally, a distance analysis between the protonated H^+ and methoxy carbonyl O showed a peak at a distance close to the H–O hydrogen bond distance (Figure 5.16 b) [196]. This leads to the postulation that the methoxy ester rigidity is primarily due to its formation of an intra-molecular hydrogen bond (IHB) with the N^+.

Additionally, the most favourable location of water molecules around cocaine was calculated to be a pocket between the same amine and carbonyl atoms.

This further leads to the hypothesis that the IHB formed is mediated by water molecules.

5.4.2.3 Cocaine diastereoisomers in MD at infinite dilution

5.4.2.3.1 Molecular Conformation

The structure of cocaine and its isomers impose a degree of structural rigidity on them in MD. With the steric constraint on the tropane motif, the carboxy-methyl and benzoyl esters remain as the only flexible groups. The EPSR results in Section 5.4.2.2 suggest an IHB confers rigidity on the methoxy ester group. This IHB has also been postulated by the MD findings of Rincon et al. [200] and DFT calculations of Zhan et al. [201].

Figure 5.17a plots the RDF of the methoxy-carbonyl O around the protonated nitrogen atom for the four cocaine diastereoisomers. In the case of cocaine, the postulation of an IHB from EPSR is supported by a sharp peak at the ideal N-H–O hydrogen bond length ($\sim$ 3.0Å) [202]. Figure 5.17 b plots the distribution of dihedral angles subtended through rotation of the C_{12}-C_{15} bond (see Figure 5.5). Similar to EPSR, the methoxy ester group is rigid with the plot exhibiting a single sharp peak at 55°. In the case of pseudococaine and allopseudococaine, the methoxy ester group is oriented away from the amine and an IHB is not possible. Their dihedral plots in Figure 5.17b correspondingly show a smaller and broader peak.

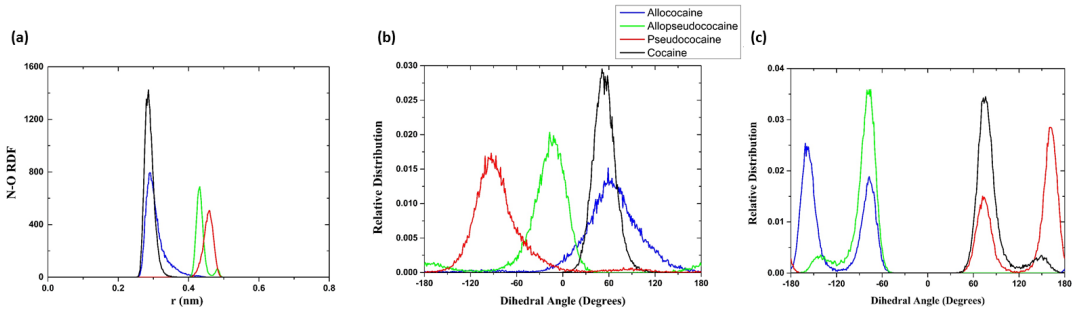


Figure 5.17 (a) RDF plots of methoxy carbonyl O around the protonated N for the four cocaine diasteroisomers. **(b)** Distribution of dihedral angles subtended by the rotation of the C_{12}-C_{15} bond for the four cocaine diasteroisomers. **(c)** Distribution of dihedral angles subtended by the rotation of the C_{10}-O_8 bond for the four cocaine diasteroisomers.

However, the RDF and dihedral angle plots for allococaine suggests an additional reason for the structural rigidity of Cocaine. Despite the methoxy-ester group oriented towards the amine, conformal rigidity is not observed to the same extent as in cocaine. Both the RDF plots in Figure 5.17a and the dihedral angle distribution in Figure 5.17b display smaller and broader peaks. Table 5.8 lists the mean and standard deviation of N-O_{20} distance for cocaine and allococaine. Cocaine has an N-O_{20} distance closer to that of an H-bond and and is more rigid with a lower standard deviation of distances than for allococaine.

Table 5.8 Comparison of the N-O_{20} bond between cocaine and allococaine.

Stereoisomer	Mean N-O_{20} distance	Std. Dev. of distance
Cocaine	2.904	0.1412
Allococaine	3.114	0.3229

These results suggest that merely the IHB is insufficient in conferring rigidity upon cocaine and that the stereochemistry of the benzoyl group is also involved. With the two ester groups oriented in the same direction of the tropane plane in cocaine, a rotation of the methoxy group can cause a '*coming together*' of the carbonyl O atoms. With an opposite benzoyl configuration in allococaine,

the methoxy-ester group is less conformally rigid. Figure 5.17c elaborates this interplay in relative rigidity of the two ester groups by plotting the distribution of the dihedral angle subtended by rotation of the O_8-C_5 bond (see Figure 5.5). In isomers with similar orientation of ester groups (cocaine and allopseudococaine), the benzoyl groups are comparatively rigid. With opposite orientations in pseudococaine and allococaine, the benzoyl groups are unhindered to alternate between two dihedral angles ($\sim 55°$ and $\sim 175°$). Yang et al. [203] observed similar results when calculating the semi-empirical potential energies of the rotation of the two ester group dihedrals. In the case of cocaine, a single deep trough in the potential energy of rotation of the two dihedrals was observed. However, a replacement of the benzoyl ester group with benzene negated the electrostatic repulsion and significantly reduced the depth of the rotational energy function.

5.4.2.3.2 Hydration

Figure 5.18 plots the results of the WMA on cocaine and displays the most favoured orientation of water around the entire molecule. In large agreement with the EPSR study, water molecules preferentially reside in a pocket between the amine and methoxy-carbonyl groups. This agreement between the two methodologies is despite the differences in solute concentration between them.

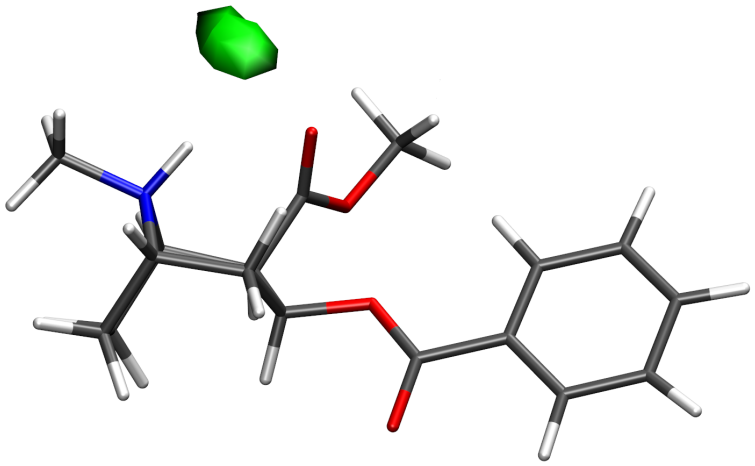


Figure 5.18 Snapshot showing the most favoured location of water molecules around the cocaine molecule from MD simulation at infinite dilution.

Figure 5.19 plots the Radial Distribution Function and the co-ordination numbers (CN) of water around the protonated amine group of the four diastereoisomers from MD at infinite dilution. All the plots in Figure 5.19a exhibit a prominent peak at $\sim$ 2.8Å - close to the experimentally observed N–H-O hydrogen bond length. The size of the peak and the co-ordination numbers in Figure 5.18b together suggest that the amine groups of pseudo- and allopseudococaine are the most hydrated followed by allococaine and cocaine.

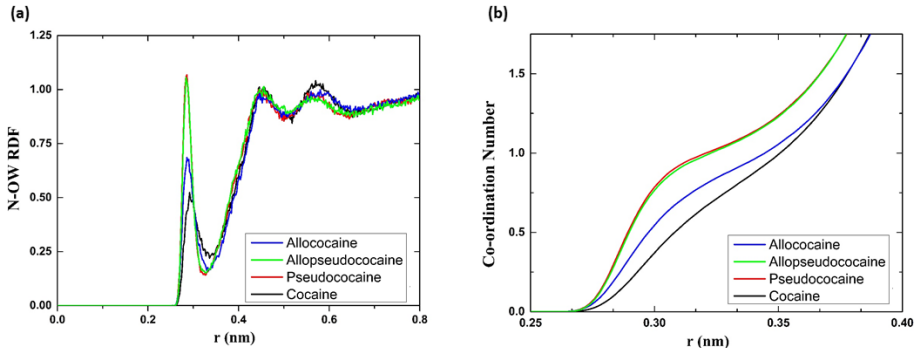


Figure 5.19 (a) RDF plots of Oxygen atom of Water around the protonated amine group of the four diastereoisomers from MD at infinite dilution. **(b)** Co-ordination numbers of the Oxygen atom of Water relative to the protonated amine group of the four diastereoisomers.

On the surface, this higher RDF and CN seem to suggest that the amine groups of the stereoisomers have greater interactions with water than Cocaine. However, the methoxy-ester group in allococaine and cocaine are oriented towards the amine group and thus make the N^+ less solvent accessible. The lower solvent accessibility allows for lesser number of sites for the binding of waters but does not necessarily imply a lower probability of finding bound water molecules. Figure 5.20 elaborates this by plotting the SDM contours of top 30% of waters around the methoxy-ester carbonyl O (O_{20} in Figure 5.5) with the positions of the N atoms added for clarity in the case of allococaine and cocaine. In the case of pseudo- and allopseudococaine, the regions of high water occupancy are broadly distributed. However, the distribution is directional in the other two isomers and

the highest occupancy is observed in the region between the N and carbonyl O atoms.

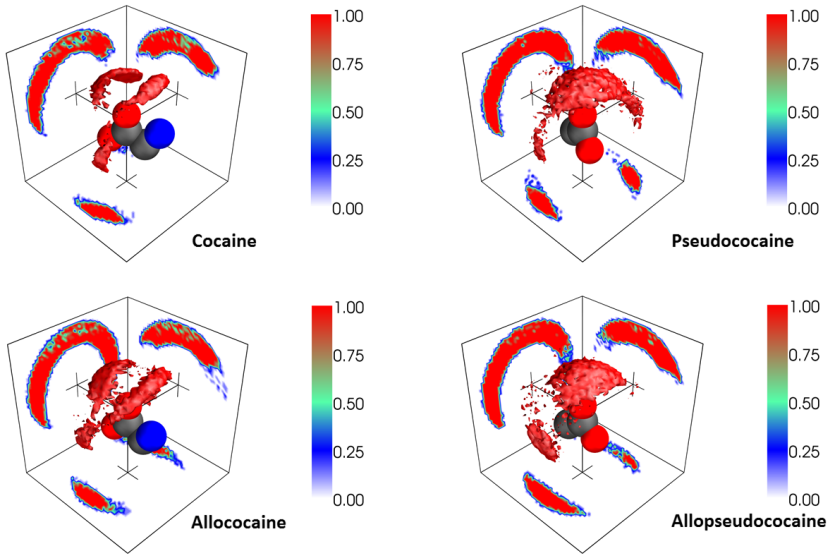


Figure 5.20 SDMs of the location of the top 30% of water molecules around the methoxy ester carbonyl O for the four stereoisomers.

5.4.2.4 Cocaine diastereoisomers in MD at concentrated solution

Figure 5.21a plots the Center-of-Mass (COM) intermolecular RDFs of the four isomers from MD simulations at a 1:100 ratio. While all stereoisomers display a small peak at 5.5 Å, the cocaine-cocaine RDF surprisingly displays a large peak at 4 Å. The prominence of the peak and the distance (4 Å) suggest the presence significant agglomeration in the MD simulations of cocaine. Figure 5.21b plots the same intermolecular RDF of cocaine across the four dilution ratios. It shows that the agglomerative behaviour is prevalent even at higher dilutions. With the three isomers possessing similar charge distribution (see Table 5.5), the agglomerative behaviour should be dependent on the unique stereochemistry of cocaine.

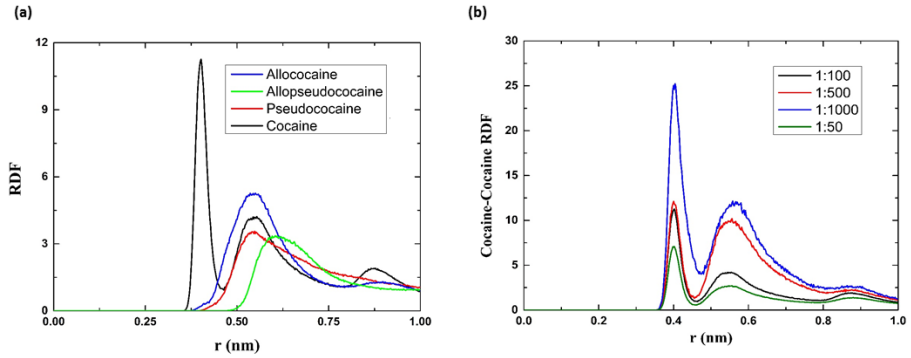


Figure 5.21 (a) COM-COM RDF of the four cocaine stereoisomers from MD at 1:500 dilution ratio. **(b)** COM-COM RDF of cocaine from MD at 1:50, 1:100, 1:500 and 1:1000 dilution ratios.

Figure 5.22a plots the SDMs of the centre-of-mass (COM) of cocaine molecules in the 1:500 dilution upto 7 Å around other cocaine molecules while 5.22b plots a similar contour for pseudococaine. The SDM in Figure 5.22a displays a single contour adjacent to the amine group and located at a region similar to the favoured water-site predicted by EPSR. In contrast, the SDM of the non-agglomerating pseudococaine has a more disperse contour. This preferential orientation of cocaine molecules on one 'side' (above the tropane plane) suggests a directional inter-molecular interaction rather than a general 'coming together'.

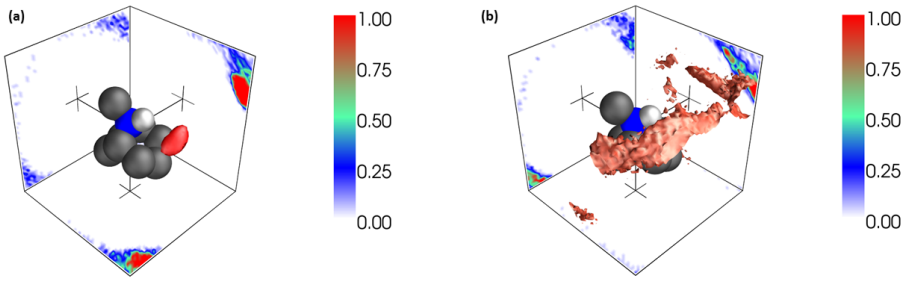


Figure 5.22 (a) SDMs of the location of the COM of the top 50% of cocaine around each other at 1:500 dilution ratio.**(b)** SDMs of the location of the COM of the top 50% of pseudococaine around each other at 1:500 dilution ratio.

Figure 5.23a is a snapshot of the most probable orientation of tropane motifs of adjacent cocaine molecules during the trajectory. Two adjacent cocaine molecules dimerize through dual inter-molecular H-bonds between the protonated amine

group and carbonyl O. This 'flipped' dimerization is consistent with the SDM in Figure 5.22a which shows a single contour for the location of cocaine molecules around each other. Figure 5.23b plots the number of cocaine agglomerates in the 1:100 system over the course of the simulation. As the simulation progresses, the number of 'single' cocaine molecules reduce and the number of cocaine molecules existing as dimers steadily increase. At around 25 ns, the system stabilizes where roughly equal number of cocaine molecules exist singly or as dimers. However, larger agglomerates (> 2) are rarely formed and are transient. Figure 5.23c plots the number of cocaine molecules that exist as dimer over the course of the simulation in the four dilution ratios. As cocaine is increasingly diluted, the number of dimers in the solution reduce and more molecules tend to reside singly.

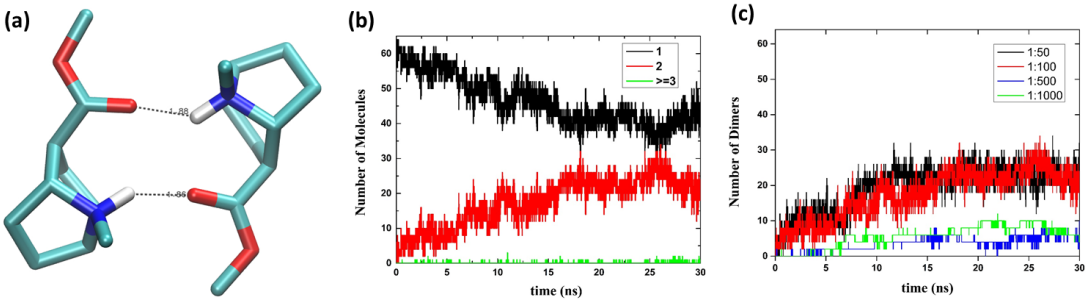


Figure 5.23 (a) Snapshot from MD describing the most probable orientation of cocaine molecules relative to each other.**(b)** Plot describing the number of singlets, dimers and tetramers of cocaine over the course of MD simulation at 1:100 dilution. **(c)** Plot describing the number of dimers of cocaine from MD simulation at various dilution ratios.

Figure 5.24 a and b investigate the disruption to the hydration structure of cocaine due to this agglomeration. Each cocaine molecule in every frame of the 1:100 dilution simulation is checked for agglomeration by calculating if its distance to the nearest cocaine COM is less than 5 Å. The water molecules around the '*agglomerated*' and '*non-agglomerated*' tropane motifs are mapped to two different grids on the origin before being histogrammed and plotted in Figures a and b respectively. In the '*non-agglomerated*' form, the hydration of cocaine is similar to that observed in EPSR and at infinite dilution MD. However, this hydration pocket is completely unoccupied by water in cocaine*s* having another cocaine

molecule within 5 Å of its COM. With Figure 5.23 predicting upto 50% of cocaines existing as dimers, the hydration shell of a large portion of cocaine molecules does not encompass the site of crystallographic water seen in cocaine-bound complexes **4XP4** [143] and **1I7Z** [144].

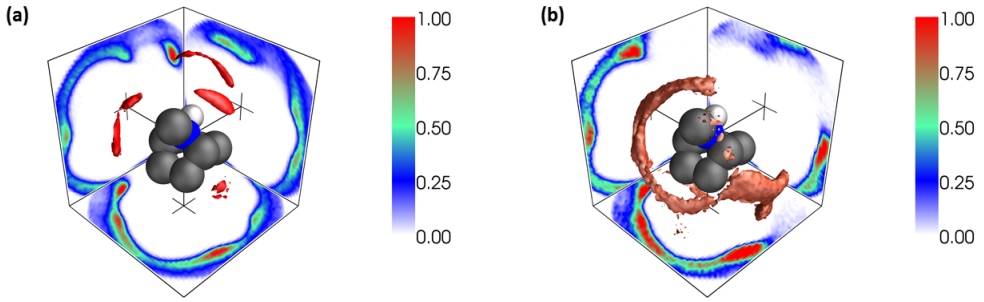


Figure 5.24 (a) SDMs of the location of water around the tropane nucleus of *agglomerated* cocaine molecules. **(b)** SDMs of the location of water around the tropane nucleus of *non-agglomerated* cocaine molecules.

5.4.3 Comparing MD and EPSR[3]

Alprazolam and cocaine present two contrasting pictures when used to compare the predictions of MD and EPSR. In the case of alprazolam, the solvation structure predicted by the two methods largely agree with each other. Conversely, the solvation of cocaine predicted by the two methods are contradictory. This dissimilarity in the case of cocaine seems to primarily result from its dimerization during MD simulations at experimental concentrations. Figure 5.25 a and b plot the Radial Distribution Functions (RDF) of water around the two atoms primarily involved in dimerization - N and O_{20}. The hydration structure of EPSR is most similar to MD simulations at infinite dilution with the hydration of the two atoms reducing at higher concentrations. Figure 5.25 c plots the inter-molecular N-O_{20} RDF from MD simulation at 1:100 ratio and EPSR. The negligible Cocaine dimerization in EPSR is apparent with the plot showing no peak at the H-bonding distance.

[3]Unpublished Data from the EPSR Cocaine and Alprazolam simulations performed by Andrew Johnston.

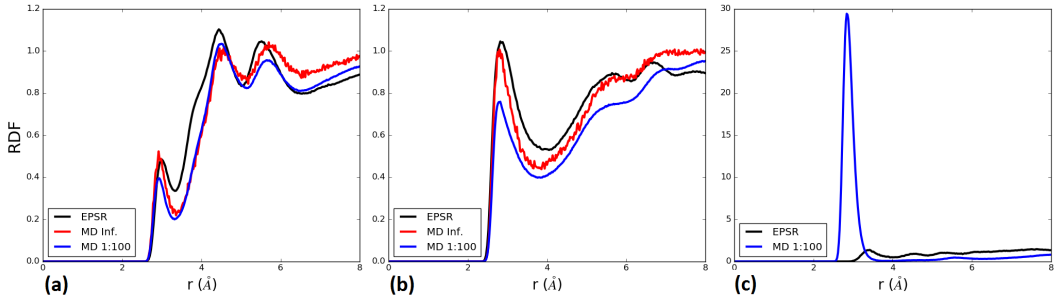


Figure 5.25 (a) RDF Plots of water oxygen atoms around the cocaine N atoms from EPSR, MD at infinite dilution and MD at 1:100 concentration ratio. **(b)** RDF plots of water oxygen atoms around the cocaine methoxy-ester carbonyl O atoms from EPSR, MD at infinite dilution and MD at 1:100 concentration ratio. **(c)** Inter-molecular RDF of the O_{20} around the protonated nitrogen atom from EPSR and MD at 1:100 concentration ratio.

Figure 5.26 compares alprazolam-alprazolam interactions in the three different simulation boxes by plotting the inter-molecular RDFs of a few salient atom sites. With an increase in MD temperature from 300 K to 323 K and the associated increase in energy, alprazolam molecules are more mobile in solution and cluster lesser. However, even at the elevated temperature, inter-solute interactions are greater in MD when compared to EPSR. This is most evident in the C6-C7 RDF plot with MD displaying a peak at $\sim$ 4Å.

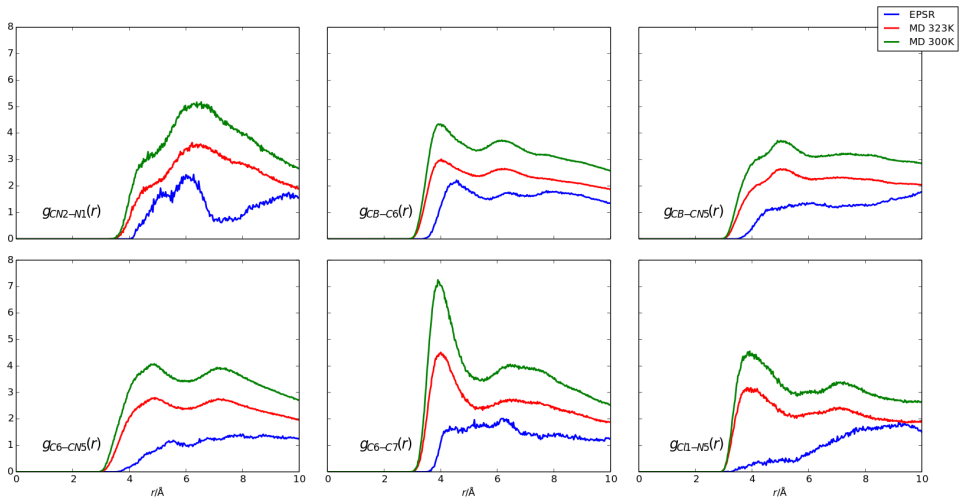


Figure 5.26 Inter-molecular RDF plots investigating alprazolam-alprazolam interactions in the three simulation boxes - EPSR (blue), MD 323K (red), MD 300K (blue).

Figure 5.27 investigates the inter-molecular interactions of alprazolam by plotting the SDM of the central diazepine rings around each other. The SDM

map indicates alprazolam-alprazolam interactions are primarily through stacking of the diazepine rings. However, the density of the contour is much lower than that plotted for cocaine-cocaine interactions in Figure 5.22a. The results suggest that while MD samples greater inter-alprazolam interaction at the concentrated ratio, the interactions are not concerted enough to disrupt solvation like in cocaine. However, this increased interaction in MD could cause the more directional water contour seen around the N2 atom in Figure 5.10k. With the stacking of diazepine rings in MD, water molecules above the plane of the ring system are extruded. The extruded water molecules then hydrate the external N2 atom. Conversely, with limited alprazolam-alprazolam interactions in EPSR, the contour of water molecules in Figure 5.10g also encompasses the region above the plane of the ring system. A combined EPSR and MD study on the hydration of caffeine by Tavagnacco et al. [204] also observed increased stacking of ring systems in MD. Similar to the current results, their study also observed increased hydration of external nitrogen atoms in MD.

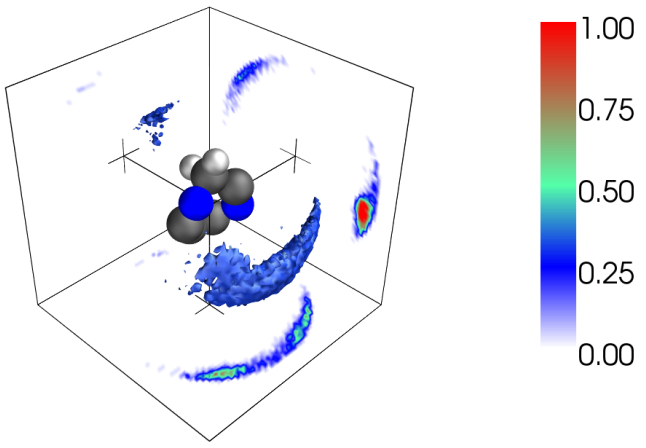


Figure 5.27 SDM plot showing the nearest-neighbour position of diazepine ring centres around other alprazolam diazepine.

5.5 Conclusion and Further Work

Neutron diffraction supported by EPSR offers an attractive methodology to validate the hydration findings of MD simulations. The preferred orientations of

water predicted through EPSR matches the locations of crystallographic waters in both alprazolam and cocaine. However, at the concentrated ratio necessitated by neutron diffraction, a direct comparison of MD and EPSR simulation boxes is not possible. EPSR underestimates inter-solute interactions in the case of both cocaine and alprazolam. In cocaine, the dimerization observed in MD disrupts the predicted hydration shell. In alprazolam however, the disruption to hydration shell caused by ring stacking is more subtle.

Together, the results of alprazolam and cocaine raise two important issues with regards to the study of hydration. Firstly, the insufficient sampling of inter-solute interactions in EPSR suggests that a direct comparison to MD can better be made at infintie dilution rather than at similar box concentrations. Experimental and computational results of a larger dataset of ligands need to be compiled to validate the trends observed.

Secondly, the results pose an interesting question with regards to the concentration at which ligand hydration should be studied. In addition to protein binding and bridging waters, ligand hydration can offer perspectives into how the molecules function *in-vivo*. These effects include bilayer translocation, competitive interactions or even solubility itself. However, whether hydration studies should consider inter-solute interactions is a question that remains unanswered.

Chapter 6

Conclusion and Future Work

6.1 Role of Ligand in Bridging Waters

Water molecules located within the binding pocket of proteins fulfil an important role in abetting interactions with the ligand. 'Conserved' water molecules form mediating interactions with both protein and ligand and in-effect act as a 'glue' between them. 'Displaced' water molecules contribute the entropy gained from their removal to the bulk to the free-energy of binding of the ligand. However, the conservation/displacement of water molecules is plausibly not completely ligand independent as previously suggested. While protein interactions do seem to play the deterministic role, the locations of all bridging water molecules overlapping with ligand hydration sites suggests that favourable interactions with the ligand is necessary. More interestingly, a simple occupancy metric of ligand-water interactions plausibly identified five ligands with tightly bound water molecules conserved across their binding to different protein structures. These water molecules are not present in the *apo* structure of the protein and can be considered 'ligand-bound'.

Further work in this regard can involve the simulation of a larger dataset of ligands together with a more intensive scoring protocol like IFST. This can validate the ligand-dependence of bridging water-sites and more importantly identify an energetic cut-off for the ligand-bound waters conserved across proteins. While only five structures were identified in this work, they represent 1% if the ∼500

analyzed. A similar percentage within the millions of structures typically analyzed by drug-screening protocols might represent a sizeable number of molecules whose ligand-bound waters can mediate interactions with the protein.

The inclusion of ligand configuration and its associated hydration shell in the prediction algorithms can significantly reduce the number of false-positive predictions with no drop in accuracy. While the current work demonstrates the methodology, it can be expanded to include the hydration shells around more 'niche' functional groups like epoxides, aziridines or peroxides. Additionally, while AutoDock Vina's empirical scoring functional is employed in this work, a cross-adaptation to other docking programs necessitates an optimization of the cut-off score for a different scoring function.

Future work can also be aimed at combining the WaterDock 2.0 protocol with drug-screens to develop a 'hydrated' docking score. With WaterDock 2.0 sensitive to changes in ligand conformation, a re-scoring of the various predicted poses to include their respective water molecules can plausibly better predictions and also allow better discrimination between binders and non-binders.

6.2 Correlating MD and EPSR

A direct comparison of the MD and EPSR simulation boxes at experimental concentration suggests a contrast in the extent of inter-solute interactions predicted. In the case of both molecules analyzed, MD comparatively predicts greater solute-solute interactions. In alprazolam, the greater stacking observed in MD only subtly affects its predicted solvation structure. However, MD predicts stable dimerization of cocaine molecules at the concentrations used for neutron diffraction and thus a complete disruption of the hydration shell observed in 'single' cocaine molecules. Further work in this regard is necessary to identify and resolve the source of the discrepancy.

Firstly, ^{1}H and ^{13}C NMR experiments on cocaine are currently in progress to validate the dimerized model of solvation predicted by MD. Preliminary ^{1}H NMR results encouragingly confirm the predictions of MD with the hydrogen atoms close to the dimerizing interface displaying a large downward shift with increasing concentration in water. In methanol conversely, an increase in concentration causes no shifting of the peaks plausibly due to amine deprotonation preventing dimerization.

A plausible reason for the reduced inter-solute interactions in EPSR might be due to the underlying MC protocol. Traditional MC simulations employing the metropolis criteria have been known to inadequately sample system configurations especially at high densities [90]. Further work can be directed towards a modification of the EPSR pathway to employ a Molecular Dynamics simulation in lieu of Monte Carlo (MC).

References

[1] Philip Ball. Water as an active constituent in cell biology. *Chemical Reviews*, 108(1):74–108, 2008.

[2] Martin Chaplin. Do we underestimate the importance of water in cell biology? *Nature Reviews Molecular Cell Biology*, 7(11):861–866, 2006.

[3] K. Tompa, M. Bokor, T. Verebelyi, and P. Tompa. Water rotation barriers on protein molecular surfaces. *Chemical Physics*, 448:15–25, 2015.

[4] Yaakov Levy and José N. Onuchic. Water Mediation in Protein Folding and Molecular Recognition. *Annual Review of Biophysics and Biomolecular Structure*, 35(1):389–415, 2006.

[5] Eric Westhof. Water : An Integral Part of Nucleic Acid Structure. *Annual Review of Biophysics and Biophysical Chemistry*, 17(1):125–144, 1988.

[6] V P Denisov, B H Jonsson, and B Halle. Hydration of denatured and molten globule proteins. *Nature Structural Biology*, 6(3):253–260, 1999.

[7] Tara Mirzadegan, Gil Benko, Slawomir Filipek, and Krzysztof Palczewski. Sequence analyses of G-protein-coupled receptors: Similarities to rhodopsin. *Biochemistry*, 42(10):2759–2767, 2003.

[8] Tetsuji Okada, Yoshinori Fujiyoshi, Maria Silow, Javier Navarro, Ehud M Landau, and Yoshinori Shichida. Functional role of internal water molecules in rhodopsin revealed by x-ray crystallography. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5982–5987, 2002.

[9] V P Denisov and B Halle. Hydrogen exchange and protein hydration: the deuteron spin relaxation dispersions of bovine pancreatic trypsin inhibitor and ubiquitin. *Journal of Molecular Biology*, 245(5):698–709, 1995.

[10] V P Denisov and B Halle. Protein hydration dynamics in aqueous solution: a comparison of bovine pancreatic trypsin inhibitor and ubiquitin by oxygen-17 spin relaxation dispersion. *Journal of Molecular Biology*, 245:682–697, 1995.

[11] Susanne Muller, Panagis Filippakopoulos, and Stefan Knapp. Bromodomains as therapeutic targets. *Expert Reviews in Molecular Medicine*, 13(29):1–21, 2011.

[12] Panagis Filippakopoulos and Stefan Knapp. Targeting bromodomains: epigenetic readers of lysine acetylation. *Nature Reviews Drug Discovery*, 13(5):337–356, 2014.

[13] Matteo Aldeghi, Alexander Heifetz, Michael J. Bodkin, Stefan Knapp, and Philip C. Biggin. Accurate calculation of the absolute free energy of binding for drug molecules. *Chemical Science*, 7(1):207–218, 2015.

[14] Yipin Lu, Renxiao Wang, Chao Yie Yang, and Shaomeng Wang. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *Journal of Chemical Information and Modeling*, 47(2):668–675, 2007.

[15] Thijs Beuming, Ramy Farid, and Woody Sherman. High-energy water sites determine peptide binding affinity and specificity of PDZ domains. *Protein Science*, 18(8):1609–1619, 2009.

[16] Benjamin Breiten, Matthew R Lockett, Woody Sherman, Shuji Fujita, Heiko Lange, Carleen M Bowers, Annie Heroux, and George M Whitesides. Water Networks Contribute to Enthalpy / Entropy Compensation in Protein-Ligand Binding. *Journal of the American Chemical Society*, 135(41):15579–15584, 2013.

[17] Geoffrey A. Holdgate, Alan Tunnicliffe, Walter H J Ward, Simon A. Weston, Gina Rosenbrock, Peter T. Barth, Ian W F Taylor, Richard A. Pauptit, and David Timms. The entropic penalty of ordered water accounts for weaker binding of the antibiotic novobiocin to a resistant mutant of DNA gyrase: A thermodynamic and crystallographic study. *Biochemistry*, 36(32):9663–9673, 1997.

[18] Jack D Dunitz. Entropic Cost of Bound Water in Crystals and Biomolecules. *Science-AAAS-Weekly Paper Edition-including Guide to Scientific Information*, 264(5159):6701–671, 1994.

[19] Caterina Barillari, Justine Taylor, Russell Viner, and Jonathan W. Essex. Classification of water molecules in protein binding sites. *Journal of the American Chemical Society*, 129(9):2577–2587, 2007.

[20] Alessio Amadasi, Francesca Spyrakis, Pietro Cozzini, Donald J. Abraham, Glen E. Kellogg, and Andrea Mozzarelli. Mapping the energetics of water-protein and water-ligand interactions with the "natural" HINT forcefield: Predictive tools for characterizing the roles of water in biomolecules. *Journal of Molecular Biology*, 358(1):289–309, 2006.

[21] Glen E. Kellogg, Simon F. Semus, and Donald J. Abraham. HINT: A new method of empirical hydrophobic field calculation for CoMFA. *Journal of Computer-Aided Molecular Design*, 5(6):545–552, 1991.

[22] Deliang L. Chen and Glen E. Kellogg. A Computational tool to optimize ligand selectivity between two similar biomacromolecular targets. *Journal of Computer-Aided Molecular Design*, 19(2):69–82, 2005.

[23] Themis Lazaridis. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory. *The Journal of Physical Chemistry B*, 102(18):3531–3541, 1998.

[24] Themis Lazaridis. Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids. *The Journal of Physical Chemistry B*, 102(18):3542–3550, 1998.

[25] Thijs Beuming, Ye Che, Robert Abel, Byungchan Kim, Veerabahu Shanmugasundaram, and Woody Sherman. Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization. *Proteins: Structure, Function and Bioinformatics*, 80(3):871–883, 2012.

[26] Schrödinger LLC. WaterMap.

[27] Alfonso T. García-Sosa, Ricardo L. Mancera, and Philip M. Dean. WaterScore: A novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *Journal of Molecular Modeling*, 9(3):172–182, 2003.

[28] M L Raymer, P C Sanschagrin, W F Punch, S Venkataraman, E D Goodman, and L A Kuhn. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *Journal of Molecular Biology*, 265(4):445–464, 1997.

[29] Gregory A. Ross, Garrett M. Morris, and Philip C. Biggin. Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLoS ONE*, 7(3):1–16, 2012.

[30] Joachim Vogt, Remo Perozzo, Alex Pautsch, Andrea Prota, Pierre Schelling, Bea Pilger, Gerd Folkers, Leonardo Scapozza, and Georg E. Schulz. Nucleoside binding site of Herpes simplex type 1 thymidine kinase analyzed by X-ray crystallography. *Proteins: Structure, Function and Genetics*, 41(4):545–553, 2000.

[31] H. Ni, C. A. Sotriffer, and J. A. McCammon. Ordered water and ligand mobility in the HIV-1 integrase-5CITEP complex: A Molecular Dynamics Study. *Journal of Medicinal Chemistry*, 44(19):3043–3047, 2001.

[32] E Chung, D Henriques, D Renzoni, M Zvelebil, J M Bradshaw, G Waksman, C V Robinson, and J E Ladbury. Mass spectrometric and thermodynamic studies reveal the role of water molecules in complexes formed between SH2 domains and tyrosyl phosphopeptides. *Structure*, 6(9):1141–1151, 1998.

[33] F A Quiocho, D K Wilson, and N K Vyas. Substrate specificity and affinity of a protein modulated by bound water molecules. *Nature*, 340(6232):404–407, 1989.

[34] Jeremy R.H. Tame, Sara H. Sleigh, Anthony J. Wilkinson, and John E. Ladbury. The role of water in sequence-independent ligand binding by an oligopeptide transporter protein. *Nature Structural Biology*, 3(12):998–1001, 1998.

[35] Bingjie Hu and Markus A Lill. Protein Pharmacophore Selection Using Hydration-Site Analysis. *Journal of Chemical Information and Modelling*, 52(4):1046–1060, 2012.

[36] Marcel L Verdonk, Gianni Chessari, Jason C Cole, Michael J Hartshorn, Christopher W Murray, J Willem M Nissink, Richard D Taylor, and Robin Taylor. Modeling Water Molecules in Protein - Ligand Docking Using GOLD. *Journal of Medicinal Chemistry*, 48(20):6504–6515, 2005.

[37] Hanzi Sun, Lifeng Zhao, Shiming Peng, and Niu Huang. Incorporating replacement free energy of binding-site waters in molecular docking. *Proteins: Structure, Function and Bioinformatics*, 82(9):1765–1776, 2014.

[38] A. Wissner, D. M. Berger, D. H. Boschelli, M. Brawner Floyd, L. M. Greenberger, B. C. Gruber, B. D. Johnson, N. Mamuya, R. Nilakantan, M. F. Reich, R. Shen, H. R. Tsou, E. Upeslacis, Yu Fen Wang, B. Wu, F. Ye, and

N. Zhang. 4-Anilino-6,7-Dialkoxyquinoline-3-Carbonitrile Inhibitors of Epidermal Growth Factor Receptor Kinase and Their Bioisosteric Relationship To the 4-Anilino-6,7-Dialkoxyquinazoline Inhibitors. *Journal of Medicinal Chemistry*, 43(17):3244–3256, 2000.

[39] James M. Chen, Simon L. Xu, Zdzislaw Wawrzak, Gregory S. Basarab, and Douglas B. Jordan. Structure-based design of potent inhibitors of scytalone dehydratase: Displacement of a water molecule from the active site. *Biochemistry*, 37(51):17735–17744, 1998.

[40] P Y Lam, P K Jadhav, C J Eyermann, C N Hodge, Y Ru, L T Bacheler, J L Meek, M J Otto, M M Rayner, and Y N Wong. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science (New York, N.Y.)*, 263(5145):380–384, 1994.

[41] K. Kinkead Reiling, Nicholas F. Endres, Deborah S. Dauber, Charles S. Craik, and Robert M. Stroud. Anisotropic dynamics of the JE-2147-HIV protease complex: Drug resistance and thermodynamic binding mode examined in a 1.09 A structure. *Biochemistry*, 41(14):4582–4594, 2002.

[42] Kristina Backbro, Seved Lowgren, Katrin Osterlund, Johnson Atepo, Torsten Unge, Johan Hulten, Nicholas M. Bonham, Wesley Schaal, Anders Karlen, and Anders Hallberg. Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor. *Journal of Medicinal Chemistry*, 40(6):898–902, 1997.

[43] Zdzislaw Wawrzak, Tatyana Sandalova, James J. Steffens, Gregory S. Basarab, Tomas Lundqvist, Ylva Lindqvist, and Douglas B. Jordan. High-resolution structures of scytalone dehydratase-inhibitor complexes crystallized at physiological pH. *Proteins*, 35(4):425–439, 1999.

[44] Renuka Kadirvelraj, B. Lachele Foley, Jane D. Dyekjær, and Robert J. Woods. Involvement of Water in Carbohydrate-Protein Binding-Concanavalin A Revisited. *Journal of the American Chemical Society*, 130(50):16933–16942, 2008.

[45] James H. Naismith and Robert A. Field. Structural basis of trimannoside recognition by concanavalin A. *Journal of Biological Chemistry*, 271(2):972–976, 1996.

[46] Paul D. Kirchhoff, Rob Brown, Scott Kahn, Marvin Waldman, and C. M. Venkatachalam. Application of structure-based focusing to the estrogen receptor. *Journal of Computational Chemistry*, 22(10):993–1003, 2001.

[47] David G. Lloyd, Alfonso T. García-Sosa, Ian L. Alberts, Nikolay P. Todorov, and Ricardo L. Mancera. The effect of tightly bound water molecules on the structural interpretation of ligand-derived pharmacophore models. *Journal of Computer-Aided Molecular Design*, 18(2):89–100, 2004.

[48] Sheng You Huang and Xiaoqin Zou. Advances and challenges in Protein-ligand docking. *International Journal of Molecular Sciences*, 11(8):3016–3034, 2010.

[49] Inbal Halperin, Buyong Ma, Haim Wolfson, and Ruth Nussinov. Principles of Docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function and Genetics*, 47(4):409–443, 2002.

[50] Yu Chian Chen. Beware of docking! *Trends in Pharmacological Sciences*, 36(2):78–95, 2015.

[51] Chris De Graaf, Pavel Pospisil, Wouter Pos, Gerd Folkers, and Nico P E Vermeulen. Binding mode prediction of cytochrome P450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *Journal of Medicinal Chemistry*, 48(7):2308–2318, 2005.

[52] Garrett M. Morris, Huey Ruth, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, 2009.

[53] Matthias Rarey, Bernd Kramer, and Thomas Lengauer. The particle concept: Placing discrete water molecules during protein- ligand docking predictions. *Proteins: Structure, Function and Genetics*, 34(1):17–28, 1999.

[54] Daniele Bellocchi, Antonio Macchiarulo, Gabriele Costantino, and Roberto Pellicciari. Docking studies on PARP-1 inhibitors: Insights into the role of a binding pocket water molecule. *Bioorganic and Medicinal Chemistry*, 13(4):1151–1157, 2005.

[55] Michael J. Hartshorn, Marcel L. Verdonk, Gianni Chessari, Suzanne C. Brewerton, W. T M Mooij, Paul N. Mortenson, and Christopher W. Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry*, 50(4):726–741, 2007.

[56] E. E. Kim, C. T. Baker, M. D. Dwyer, M. A. Murcko, B. G. Rao, R. D. Tung, and M. A. Navia. Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *Journal of the American Chemical Society*, 117(3):1181–1182, 1995.

[57] S. Maignan, J. P. Guilloteau, S. Pouzieux, Y. M. Choi-Sledeski, M. R. Becker, S. I. Klein, W. R. Ewing, H. W. Pauls, A. P. Spada, and V. Mikol. Crystal structures of human factor Xa complexed with potent inhibitors. *Journal of Medicinal Chemistry*, 43(17):3226–3232, 2000.

[58] Oliviero Carugo and Domenico Bordo. How many water molecules can be detected by protein crystallography? *Acta Crystallographica Section D: Biological Crystallography*, 55(2):479–483, 1999.

[59] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[60] B P Hills, S F Takacs, and P S Belton. The effects of proteins on the proton N.M.R. transverse relaxation times of water. *Molecular Physics*, 67(4):903–918, 1989.

[61] Mette A. Lie, René Thomsen, Christian N S Pedersen, Birgit Schiøtt, and Mikael H. Christensen. Molecular docking with ligand attached water molecules. *Journal of Chemical Information and Modeling*, 51(4):909–917, 2011.

[62] G. Ratnavali, N. Devi, K. Sri, J. Raju, B. Sirisha, and R. Kavitha. An attempt to screen top colorectal cancer drugs by using Molegro Virtual Docker. *Annals of Biological Research*, 2(1):114–126, 2011.

[63] Stefano Forli and Arthur J. Olson. A force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *Journal of Medicinal Chemistry*, 55(2):623–638, 2012.

[64] Gordon Lemmon and Jens Meiler. Towards Ligand Docking Including Explicit Interface Water Molecules. *PLoS ONE*, 8(6):1–12, 2013.

[65] Jens Meiler and David Baker. ROSETTALIGAND: Protein–Small Molecule Docking with Full Side-Chain Flexibility. *Proteins: Structure, Function and Bioinformatics*, 65(3):538–548, 2006.

[66] James B. Dunbar, Richard D. Smith, Kelly L. Damm-Ganamet, Aqeel Ahmed, Emilio Xavier Esposito, James Delproposto, Krishnapriya Chinnaswamy, You Na Kang, Ginger Kubish, Jason E. Gestwicki, Jeanne A. Stuckey, and Heather A. Carlson. CSAR data set release 2012: Ligands, affinities, complexes, and docking decoys. *Journal of Chemical Information and Modeling*, 53(8):1842–1852, 2013.

[67] Richard H. Henchman and J. Andrew McCammon. Extracting hydration sites around proteins from explicit water simulations. *Journal of Computational Chemistry*, 23(9):861–869, 2002.

[68] Arina Afanasyeva, Sergey Izmailov, Michel Grigoriev, and Michael Petukhov. AquaBridge: A novel method for systematic search of structural water molecules within the protein active sites. *Journal of Computational Chemistry*, 36(26):1973–1977, 2015.

[69] Piotr Setny and Martin Zacharias. Hydration in Discrete Water. A mean field, cellular automata based approach to calculating hydration free energies. *Journal of Physical Chemistry B*, 114(26):8667–8675, 2010.

[70] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: An online force field. *Nucleic Acids Research*, 33(SUPPL. 2):382–388, 2005.

[71] Joost W H Schymkowitz, Frederic Rousseau, Ivo C Martins, Jesper Ferkinghoff-Borg, Francois Stricher, and Luis Serrano. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29):10147–10152, 2005.

[72] Oleg Trott and Arthur J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010.

[73] W R Pitt and J M Goodfellow. Modelling of solvent positions around polar groups in proteins. *Protein Engineering*, 4(5):531–537, 1991.

[74] Gianluca Rossato, Beat Ernst, Angelo Vedani, and Martin Smieško. AcquaAlta: A directional approach to the solvation of ligand-protein complexes. *Journal of Chemical Information and Modeling*, 51(8):1867–1881, 2011.

[75] Frank H. Allen. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallographica Section B: Structural Science*, 58(3):380–388, 2002.

[76] Firdaus Samsudin, J L Parker, Mark S. P. Sansom, Simon Newstead, and Philip W Fowler. Accurate Prediction of Ligand Affinities for a Peptide Transporter. *Cell Chemical Biology*, 23(2):299–309, 2015.

[77] Magnus Lundborg and Erik Lindahl. Automatic GROMACS topology generation and comparisons of force fields for solvation free energy calculations. *The journal of physical chemistry. B*, 119(3):810–23, 2015.

[78] Alexander D Jr MacKerell, D Bashford, M Bellott, R L Dunbrack, J D Evanseck, M J Field, S Fischer, J Gao, H Guo, S Ha, D Joseph-McCarthy, L Kuchnir, K Kuczera, F T Lau, C Mattos, S Michnick, T Ngo, D T Nguyen, B Prodhom, W E Reiher, Benoît Roux, M Schlenkrich, Jeremy C Smith, R Stote, J Straub, M Watanabe, J Wiórkiewicz-Kuczera, D Yin, and Martin Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 102(18):3586–616, 1998.

[79] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function and Bioinformatics*, 78(8):1950–1958, 2010.

[80] WL Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[81] J M Wang, R M Wolf, J W Caldwell, P A Kollman, and D A Case. Development and Testing of a General Amber Force Field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[82] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics*, 79(2):926–935, 1983.

[83] Bhanita Sharma and Sandip Paul. Understanding the Role of Temperature Change and the Presence of NaCl Salts on Caffeine Aggregation in Aqueous Solution: From Structural and Thermodynamics Point of View. *Journal of Physical Chemistry B*, 119(21):6421–32, 2015.

[84] Richard J. Gillams, Jon V. Busto, Sebastian Busch, Felix M. Goñi, Christian D. Lorenz, and Sylvia E. McLain. Solvation and hydration of the ceramide headgroup in a non-polar solution. *Journal of Physical Chemistry B*, 119(1):128–139, 2015.

[85] A.J. Johnston, Y.R. Zhang, S. Busch, L.C. Pardo, S. Imberti, and S.E. McLain. Amphipathic Solvation of Indole: Implications for the Role of Tryptophan in Membrane Proteins. *Journal of Physical Chemistry B*, 119(19):5979–5987, 2015.

[86] J. J. Towey, A. K. Soper, and L. Dougan. Preference for isolated water molecules in a concentrated glycerol-water mixture. *Journal of Physical Chemistry B*, 115(24):7799–7807, 2011.

[87] Alan K. Soper. Computer simulation as a tool for the interpretation of total scattering data from glasses and liquids. *Molecular Simulation*, 38(14-15):1171–1185, 2012.

[88] Sebastian Busch, Christian D Lorenz, Jonathan Taylor, Luis Carlos Pardo, and Sylvia E McLain. Short-range interactions of concentrated proline in aqueous solution. *Journal of Physical Chemistry B*, 118(49):14267–14277, 2014.

[89] Andrew Johnston, Sebastian Busch, Luis Carlos Pardo, Samantha Kim Callear, Philip Biggin, and Sylvia McLain. On the Atomic Structure of Cocaine in Solution. *Physical Chemistry Chemical Physics*, 4(8):1166–1169, 2015.

[90] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Academic Press, San Diego, 2 edition, 2001.

[91] Jay W. Ponder and David A. Case. Force fields for protein simulation. In Valerie Daggett, editor, *Advances in Protein Chemistry*, pages 27–85. Academic Press, Seattle, 2003.

[92] R. W. Hockney, S. P. Goel, and J. W. Eastwood. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2):148–158, 1974.

[93] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *Journal of Chemical Physics*, 98(12):10089, 1993.

[94] Björn Linse and Per Linse. Tuning the smooth particle mesh Ewald sum: Application on ionic solutions and dipolar fluids. *Journal of Chemical Physics*, 141(18):184114, 2014.

[95] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics*, 81(8):3684, 1984.

[96] H J C Berendsen, J R Grigera, and T P Straatsma. The Missing Term in Effective Pair Potentials. *Journal of Physical Chemistry*, 91(24):6269–6271, 1987.

[97] Attila G. Csaszar, Gabor Czako, Tibor Furtenbacher, Jonathan Tennyson, Viktor Szalay, Sergei V. Shirin, Nikolai F. Zobov, and Oleg L. Polyansky. On equilibrium structures of the water molecule. *Journal of Chemical Physics*, 122(21):214305, 2005.

[98] Chris Oostenbrink, Alessandra Villa, Alan E. Mark, and Wilfred F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry*, 25(13):1656–1676, 2004.

[99] V Zoete, MA Cuendet, A Grosdidier, and O Michielin. SwissParam: A Fast Force Field Generation Tool for Small Organic Molecules. *Journal of Computational Chemistry*, 32(11):2359–2368, 2011.

[100] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins: Structure, Function and Bioinformatics*, 65(3):712–725, 2006.

[101] Christopher I Bayly, Piotr Cieplak, Wendy Cornell, and Peter A Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *Journal of Physical Chemistry*, 97(40):10269–10280, 1993.

[102] Araz Jakalian, Bruce L Bush, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of Computational Chemistry*, 21(2):132–146, 2000.

[103] Araz Jakalian, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry*, 23(16):1623–1641, 2002.

[104] Brent H. Besler, Kenneth M. Merz, and Peter A. Kollman. Atomic Charges Derived from Semiempirical Methods. *Journal of Computational Chemistry*, 11(4):431–439, 1990.

[105] Joey W. Storer, David J. Giesen, Christopher J. Cramer, and Donald J. Truhlar. Class IV charge models: A new semiempirical approach in quantum chemistry. *Journal of Computer-Aided Molecular Design*, 9(1):87–110, 1995.

[106] Joakim P M Jambeck, Francesca Mocci, Alexander P. Lyubartsev, and Aatto Laaksonen. Partial atomic charges and their impact on the free energy of solvation. *Journal of Computational Chemistry*, 34(3):187–197, 2013.

[107] D Shivakumar, Y Q Deng, and B Roux. Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model. *Journal of Chemical Theory and Computation*, 5(4):919–930, 2009.

[108] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247–260, 2006.

[109] David A Case, Thomas E Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–88, 2005.

[110] Waltraud Huyer and Arnold Neumaier. Global Optimization by Multilevel Coordinate Search. *Journal of Global Optimization*, 14(4):331–355, 1999.

[111] Riccardo Baron, Piotr Setny, and J. Andrew McCammon. Water in cavity-ligand recognition. *Journal of the American Chemical Society*, 132(34):12091–12097, 2010.

[112] Riccardo Baron, Piotr Setny, and Francesco Paesani. Water Structure , Dynamics , and Spectral Signatures : Changes Upon Model Cavity Ligand Recognition. *Journal of Physical Chemistry A*, 116(46):13774–13780, 2012.

[113] P. W. Snyder, J. Mecinovic, D. T. Moustakas, S. W. Thomas, M. Harder, E. T. Mack, M. R. Lockett, A Heroux, W. Sherman, and G. M. Whitesides. Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. *Proceedings of the National Academy of Sciences*, 108(44):17889–17894, 2011.

[114] John E. Ladbury. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology*, 3(12):973–980, 1996.

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

[116] Alan W Sousa Da Silva and Wim F Vranken. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Research Notes*, 5:367–374, 2012.

[117] In Suk Joung and Thomas E. Cheatham. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *Journal of Physical Chemistry B*, 112(30):9020–9041, 2008.

[118] M. Parrinello. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182, 1981.

[119] Naveen Michaud-Agrawal, Elizabeth J Denning, Thomas B Woolf, and Oliver Beckstein. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, 32(10):2319–2327, 2011.

[120] Laurie J Heyer, Semyon Kruglyak, and Shibu Yooseph. Exploring Expression Data : Identification and Analysis of Coexpressed Genes Exploring Expression Data : Identification and Analysis of Coexpressed Genes. *Genome Research*, 9(11):1106–1115, 1999.

[121] Daniel J. Sindhikara, Norio Yoshida, and Fumio Hirata. Placevent: An algorithm for prediction of explicit solvent atom distribution-Application to HIV-1 protease and F-ATP synthase. *Journal of Computational Chemistry*, 33(18):1536–1543, 2012.

[122] Bingjie Hu and Markus A. Lill. WATsite: Hydration site prediction program with PyMOL interface. *Journal of Computational Chemistry*, 35(16):1255–1260, 2014.

[123] Schrödinger LLC. The PyMOL Molecular Graphics System, Version 1.7.4.

[124] Axel T. Brunger. Assessment of phase accuracy by cross validation: the free R value. Methods and applications. *Acta Crystallographica Section D: Biological Crystallography*, 49(1):24–36, 1993.

[125] Axel T. Brünger. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(6359):472–475, 1992.

[126] Niu Huang and Brian K. Shoichet. Exploiting ordered waters in molecular docking. *Journal of Medicinal Chemistry*, 51(16):4862–4865, 2008.

[127] RDKiT: Open-source cheminformatics.

[128] P. Ertl, B. Rohde, and P. Selzer. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20):3714–3717, 2000.

[129] Panagis Filippakopoulos, Sarah Picaud, Oleg Fedorov, Marco Keller, Matthias Wrobel, Olaf Morgenstern, Franz Bracher, and Stefan Knapp. Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family. *Bioorganic and Medicinal Chemistry*, 20(6):1878–1886, 2012.

[130] Maureen Verhaest, Willem Lammens, Katrien Le Roy, Camiel J. De Ranter, André Van Laere, Anja Rabijns, and Wim Van Den Ende. Insights into the fine architecture of the active site of chicory fructan 1-exohydrolase: 1-Kestose as substrate vs sucrose as inhibitor. *New Phytologist*, 174(1):90–100, 2007.

[131] T Zhou, S Georgeon, R Moser, D J Moore, A Caflisch, and O Hantschel. Specificity and mechanism-of-action of the JAK2 tyrosine kinase inhibitors ruxolitinib and SAR302503 (TG101348). *Leukemia*, 28(2):404–407, 2014.

[132] T L Lasho, A Tefferi, J D Hood, S Verstovsek, D G Gilliland, and A Pardanani. TG101348, a JAK2-selective antagonist, inhibits primary hematopoietic cells derived from myeloproliferative disorder patients with JAK2V617F, MPLW515K or JAK2 exon 12 mutations as well as mutation negative patients. *Leukemia*, 22(9):1790–1792, 2008.

[133] Animesh Pardanani, Jason R. Gotlib, Catriona Jamieson, Jorge E. Cortes, Moshe Talpaz, Richard M. Stone, Michael H. Silverman, D. Gary Gilliland, Jolene Shorr, and Ayalew Tefferi. Safety and efficacy of TG101348, a selective JAK2 inhibitor, in myelofibrosis. *Journal of Clinical Oncology*, 29(7):789–796, 2011.

[134] Stuart W J Ember, Jin Yi Zhu, Sanne H. Olesen, Mathew P. Martin, Andreas Becker, Norbert Berndt, Gunda I. Georg, and Ernst Schonbrunn. Acetyllysine binding site of bromodomain-containing protein 4 (BRD4) interacts with diverse kinase inhibitors. *ACS Chemical Biology*, 9(5):1160–1171, 2014.

[135] S. Munshi, Z. Chen, Y. Yan, D. B. Olsen, H. B. Schock, B. B. Galvin, B. Dorsey, and L.C. Kuo. An alternate binding site for the P1-P3 group of a class of potent HIV-1 protease inhibitors as a result of concerted structural change in the 80s loop of the protease. *Acta Crystallographica Section D*, 56(1):381–388, 2000.

[136] Fengling Liu, Peter I. Boross, Yuan-Fang Wang, Jozsef Tozser, John M. Louis, Robert W. Harrison, and Irene T. Weber. Kinetic, Stability, and Structural Changes in High-resolution Crystal Structures of HIV-1 Protease with Drug-resistant Mutations L24I, I50V, and G73S. *Journal of Molecular Biology*, 354(4):789–800, 2005.

[137] Maren Kuhnert, Holger Steuber, and Wibke E. Diederich. Structural basis for HTLV-1 protease inhibition by the HIV-1 protease inhibitor indinavir. *Journal of Medicinal Chemistry*, 57(14):6266–6272, 2014.

[138] Laurent Rivier. Analysis of Alkaloids in Leaves of Cultivated Erythroxylum and Characterization of Alkaline substances used during Coca Chewing. *Journal of Ethnopharmacology*, 3:313–335, 1981.

[139] Mauri Lounasmaa. The Tropane Alkaloids. In Arnold Brossi, editor, *The Alkaloids*, chapter 1, pages 1–81. Academic Press, San Diego, 1988.

[140] Laurent Karila, Rim Zarmdini, Aymeric Petit, Geneviève Lafaye, William Lowenstein, and Michel Reynaud. Addiction à la cocaïne: Données actuelles pour le clinicien. *Presse Medicale*, 43(1):9–17, 2014.

[141] Nicholas J Connors and Robert S Hoffman. Experimental treatments for cocaine toxicity: a difficult transition to the bedside. *Journal of Pharmacology and Experimental Therapeutics*, 347(2):251–7, 2013.

[142] Janice L. Zimmerman. Cocaine Intoxication. *Critical Care Clinics*, 28(4):517–526, 2012.

[143] Kevin H. Wang, Aravind Penmatsa, and Eric Gouaux. Neurotransmitter and psychostimulant recognition by the dopamine transporter. *Nature*, 521(7552):322–327, 2015.

[144] N A Larsen, B Zhou, A Heine, P Wirsching, K D Janda, and I A Wilson. Crystal Structure of a Cocaine-binding Antibody. *Journal of Molecular Biology*, 311(1):9–15, 2001.

[145] C. A. Beevers, T. R. R. McDonald, J. H. Robertson, and F. Stern. The crystal structure of sucrose. *Acta Crystallographica*, 5(5):689–690, 1952.

[146] T.C. Jin, F. Guo, Y.T. Wang, A. Howard, and Y.Z Zhang. Unpublished Data.

[147] Yu Chen, Julien Delmas, Jacques Sirot, Brian Shoichet, and Richard Bonnet. Atomic resolution structures of CTX-M β-lactamases: Extended spectrum activities from increased mobility and decreased stability. *Journal of Molecular Biology*, 348(2):349–362, 2005.

[148] Pietro Cozzini, Anna Fornabaio, Micaela Marabotti, Glen E. Abraham, Donald J. Kellogg, and Andrea Mozzarelli. Free Energy of Ligand Binding to Protein: Evaluation of the Contribution of Water Molecules by Computational Methods. *Current Medicinal Chemistry*, 11(23):3093–3118, 2004.

[149] Fredrik Österberg, Garrett M. Morris, Michel F. Sanner, Arthur J. Olson, and David S. Goodsell. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins: Structure, Function and Genetics*, 46(1):34–40, 2002.

[150] M Pastor, G Cruciani, and K A Watson. A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure-activity relationship analysis. *Journal of Medicinal Chemistry*, 40(25):4089–102, 1997.

[151] Abrar Hussain, James L. Melville, and Jonathan D. Hirst. Molecular docking and QSAR of aplyronine A and analogues: Potent inhibitors of actin. *Journal of Computer-Aided Molecular Design*, 24(1):1–15, 2010.

[152] Mutasem O. Taha, Maha Habash, Zeina Al-Hadidi, Amal Al-Bakri, Khaled Younis, and Suhaib Sisan. Docking-based comparative intermolecular contacts analysis as new 3-D QSAR concept for validating docking studies and in silico screening: NMT and GP inhibitors as case studies. *Journal of Chemical Information and Modeling*, 51(3):647–669, 2011.

[153] Richard A. Friesner, Robert B. Murphy, Matthew P. Repasky, Leah L. Frye, Jeremy R. Greenwood, Thomas A. Halgren, Paul C. Sanschagrin, and Daniel T. Mainz. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21):6177–6196, 2006.

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

[155] R. Becker and S. Panchanadeeswaran. Crystal Rotations Represented as Rodrigues Vectors. *Textures and Microstructures*, 10(3):167–194, 1989.

[156] Scott T. Leutenegger, Mario A. Lopez, and Jeffrey Edgington. STR: A Simple and Efficient Algorithm for R-Tree Packing. In *13th International Conference on Data Engineering*, pages 497–506, 1997.

[157] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman J C Berendsen. GROMACS: fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718, dec 2005.

[158] N. H. Rhys, A. K. Soper, and L. Dougan. The hydrogen-bonding ability of the amino acid glutamine revealed by neutron diffraction experiments. *Journal of Physical Chemistry B*, 116(45):13308–13319, 2012.

[159] VF Varley F. Sears. Neutron scattering lengths and cross sections. *Neutron News*, 3(3):26–37, 1992.

[160] Krawitz Aaron D. *Introduction to Diffraction in Material Science and Engineering*. Wiley-Interscience, New York, 2001.

[161] Alan K. Soper. Computer simulation as a tool for the interpretation of total scattering data from glasses and liquids. *Molecular Simulation*, 38(14-15):1171–1185, 2012.

[162] Camilla Haw and Jean Stubbs. Benzodiazepines–a necessary evil? A survey of prescribing at a specialist UK psychiatric hospital. *Journal of Psychopharmacology*, 21(6):645–649, 2007.

[163] J. Riss, J. Cloyd, J. Gates, and S. Collins. Benzodiazepines in epilepsy: Pharmacology and pharmacokinetics. *Acta Neurologica Scandinavica*, 118(2):69–86, 2008.

[164] A M Holbrook, R Crowther, A Lotter, C Cheng, and D King. Meta-analysis of benzodiazepine use in the treatment of insomnia. *Canadian Medical Association Journal*, 162(2):225–233, 2000.

[165] F Petty, M H Trivedi, M Fulton, and A J Rush. Benzodiazepines as antidepressants: does GABA play a role in depression? *Biological Psychiatry*, 38(9):578–591, 1995.

[166] W Hunkeler, H Mohler, L Pieri, P Polc, E P Bonetti, R Cumin, R Schaffner, and W Haefely. Selective antagonists of benzodiazepines, 1981.

[167] Panagis Filippakopoulos, Jun Qi, Sarah Picaud, Yao Shen, William B Smith, Oleg Fedorov, Elizabeth M Morse, Tracey Keates, Tyler T Hickman, Ildiko Felletar, Martin Philpott, Shonagh Munro, Michael R McKeown, Yuchuan Wang, Amanda L Christie, Nathan West, Michael J Cameron, Brian Schwartz, Tom D Heightman, Nicholas La Thangue, Christopher A French, Olaf Wiest, Andrew L Kung, Stefan Knapp, and James E Bradner. Selective inhibition of BET bromodomains. *Nature*, 468(7327):1067–1073, 2010.

[168] Eleonora Da Pozzo, Chiara Giacomelli, Elisabetta Barresi, Barbara Costa, Sabrina Taliani, Federico Da Settimo Passetti, and Claudia Martini. Targeting the 18-kDa translocator protein: recent perspectives for neuroprotection. *Biochemical Society Transactions*, 43(4):559–565, 2015.

[169] Neel A. Shah, Marie A. Abate, Michael J. Smith, James A. Kaplan, James C. Kraner, and David J. Clay. Characteristics of Alprazolam-Related Deaths Compiled by a Centralized State Medical Examiner. *American Journal on Addictions*, 21:S27–S34, 2012.

[170] Geoffrey K. Isbister, Luke O'Regan, David Sibbritt, and Ian M. Whyte. Alprazolam is relatively more toxic than other benzodiazepines in overdose. *British Journal of Clinical Pharmacology*, 58(1):88–95, 2004.

[171] Virginia D. Schmith, Beth Piraino, Randall B. Smith, and Patricia D. Kroboth. Alprazolam in End-Stage Renal Disease: I. Pharmacokinetics. *Journal of Clinical Pharmacology*, 31(6):571–579, 1991.

[172] Satwata Maitra, Biswarup Saha, Chitta Ranjan Santra, Ananda Mukherjee, Sumit Goswami, Palash Kumar Chanda, and Parimal Karmakar. Alprazolam induced conformational change in hemoglobin. *International Journal of Biological Macromolecules*, 41(1):23–29, 2007.

[173] Moumita Sarkar, Shiv Shankar Paul, and Kalyan K. Mukherjea. Interaction of bovine serum albumin with a psychotropic drug alprazolam: Physicochemical, photophysical and molecular docking studies. *Journal of Luminescence*, 142:220–230, 2013.

[174] Faeze Dangkoob, Mohmmad Reza Housaindokht, Ahmad Asoodeh, Omid Rajabi, Zeinab Rouhbakhsh Zaeri, and Asma Verdian Doghaei. Spectroscopic and molecular modeling study on the separate and simultaneous bindings of alprazolam and fluoxetine hydrochloride to human serum albumin (HSA): With the aim of the drug interactions probing. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 137:1106–1119, 2015.

[175] Biswarup Saha, Ananda Mukherjee, Chitta Ranjan Santra, Atiskumar Chattopadhyay, Amar Nath Ghosh, Utpal Choudhuri, and Parimal Karmakar. Alprazolam Intercalates into DNA. *Journal of Biomolecular Structure and Dynamics*, 26(4):421–429, 2009.

[176] M Novák, C A Salemink, and I Khan. Biological activity of the alkaloids of Erythroxylum coca and Erythroxylum novogranatense. *Journal of Ethnopharmacology*, 10(3):261–274, 1984.

[177] Arnold M. Washton and Mark S. Gold. *Cocaine: A clinician's handbook*. Guilford Press, New York, 1987.

[178] Michael E. O'Leary and Jules C. Hancox. Role of voltage-gated sodium, potassium and calcium channels in the development of cocaine-associated cardiac arrhythmias. *British Journal of Clinical Pharmacology*, 69(5):427–442, 2010.

[179] S N Wright, S Y Wang, Y F Xiao, and G K Wang. State-Dependent Cocaine Block of Sodium Channel Isoforms, Chimeras, and Channels Coexpressed with the B1 Subunit. *Biophysical Journal*, 76(1):233–245, 1999.

[180] Harry A. Fozzard, Michael F. Sheets, and Dorothy A. Hanck. The sodium channel as a target for local anesthetic drugs. *Frontiers in Pharmacology*, 2(1):1–6, 2011.

[181] Y A Ruetsch, T Böni, and A Borgeat. From cocaine to ropivacaine: the history of local anesthetic drugs. *Current topics in medicinal chemistry*, 1(3):175–82, 2001.

[182] Rong Chen, Michael R Tilley, Hua Wei, Fuwen Zhou, Fu-Ming Zhou, San Ching, Ning Quan, Robert L Stephens, Erik R Hill, Timothy Nottoli, Dawn D Han, and Howard H Gu. Abolished cocaine reward in mice with a cocaine-insensitive dopamine transporter. *PNAS*, 103(24):9333–9338, 2006.

[183] Xiaoqin Huang, Howard H Gu, and Chang-Guo Zhan. Mechanism for Cocaine Blocking the Transport of Dopamine: Insights from Molecular Modeling and Dynamics Simulations. *Journal of Physical Chemistry B*, 113(45):15057–66, 2009.

[184] E. J. Gabe and W. H. Barnes. The Crystal and Molecular structure of l-Cocaine Hydrochloride. *Acta Crystallographica*, 16(8):796–801, 1963.

[185] Hugo O Villar and Gilda H Loew. A conformational study of cocaine and its diastereomers. *Journal of Computational Chemistry*, 11(9):1111–1118, 1990.

[186] A. K. Soper. *GudrunN and GudrunX: Programs for correcting raw neutron and X-ray diffraction data to differential scattering cross section*. Rutherford Appleton Laboratory, STFC, UK, 2011.

[187] Evan E. Bolton, Jie Chen, Sunghwan Kim, Lianyi Han, Siqian He, Wenyao Shi, Vahan Simonyan, Yan Sun, Paul a. Thiessen, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. PubChem3D: A new resource for scientists. *Journal of Cheminformatics*, 3(9):1–15, 2011.

[188] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33–47, 2011.

[189] Luis Carlos Pardo, Muriel Rovira-Esteva, Josep Lluis Tamarit, Nestor Veglio, Francisco Javier Bermejo, and Gabriel Julio Cuello. A Procedure to Quantify the Short Range Order of Disordered Phases. In Sylwester Rzoska, Aleksandra Drozd-Rzoska, and Victor Mazur, editors, *Metastable Systems under Pressure*, pages 79–91. Springer Netherlands, 2010.

[190] Pu Liu, Dimitris K. Agrafiotis, and Douglas L. Theobald. Fast Determination of the Optimal Rotational Matrix for Macromolecular Superpositions. *Journal of Computational Chemistry*, 31(7):1561–1563, 2010.

[191] J. D. Hunter. Matplotlib: A 2D graphic environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[192] Prabhu Ramachandran and Gael Varoquaux. Mayavi: 3D visualization of scientific data. *Computing in Science and Engineering*, 13(2):40–51, 2011.

[193] A. K. Soper. Inelasticity corrections for time-of-flight and fixed wavelength neutron diffraction experiments. *Molecular Physics*, 107(16):1667–1684, 2009.

[194] B Jacrot. The study of biological structures by neutron scattering from solution. *Reports on Progress in Physics*, 39(10):911–953, 1976.

[195] S J Perkins. Structural studies of proteins by high-flux X-ray and neutron solution scattering. *Biochemical Journal*, 254(2):313–327, 1988.

[196] B Raghavendra, Pankaj K Mandal, and E Arunan. Ab initio and AIM theoretical analysis of hydrogen-bond radius of HD (D = F, Cl, Br, CN, HO, HS and CCH) donors and some acceptors. *Physical Chemistry Chemical Physics*, 8(45):5276–5286, 2006.

[197] Huafeng Xu and B. J. Berne. Hydrogen-bond kinetics in the solvation shell of a polypeptide. *Journal of Physical Chemistry B*, 105(48):11929–11932, 2001.

[198] Martin C. Stumpe and Helmut Grubmüller. Interaction of urea with amino acids: Implications for urea-induced protein denaturation. *Journal of the American Chemical Society*, 129(51):16126–16131, 2007.

[199] Nicola Steinke, Richard J. Gillams, Luis Carlos Pardo, Christian D. Lorenz, and Sylvia E. McLain. Atomic scale insights into urea–peptide interactions in solution. *Physical Chemistry Chemical Physics*, 18(5):10–14, 2016.

[200] David A Rincón, Maria Natália Dias Soeiro Cordeiro, and Ricardo a Mosquera. On the Electronic Structure of Cocaine and its Metabolites. *Journal of Physical Chemistry A*, 113(50):13937–13942, 2009.

[201] Chang-Guo Zhan, Shi-Xian Deng, Jaime G Skiba, Beth A Hayes, Sarah M Tschampel, George C Shields, and Donald W Landry. First-Principle studies of Intermolecular and Intramolecular Catalysis of Protonated Cocaine. *Journal of Computational Chemistry*, 26(10):980–986, 2005.

[202] S. C. Wallwork. Hydrogen-bond radii. *Acta Crystallographica*, 15(8):758–759, 1962.

[203] Biao Yang, Jeremy Wright, Mohyee E Eldefrawi, Sovitj Pou, and Alexander D Mackerell. Conformational, Aqueous Solvation, and pKa Contributions to the Binding and Activity of Cocaine, WIN 32 065-2, and the WIN Vinyl Analog. *Journal of the American Chemical Society*, 116(9):8722–8732, 1994.

[204] L. Tavagnacco, J. W. Brady, F. Bruni, S. Callear, M. A. Ricci, M. L. Saboungi, and A. Cesàro. Hydration of Caffeine at High Temperature by Neutron Scattering and Simulation Studies. *Journal of Physical Chemistry B*, 119(42):13294–13301, 2015.