Computational Modeling of Protein Structure, Function, and Binding Hotspots

by

Sarah E. Graham

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biophysics)
in the University of Michigan
2017

Doctoral Committee:

Professor Heather A. Carlson, Chair
Professor Charles L. Brooks III
Assistant Professor Tomasz Cierpicki
Professor Kevin J. Kubarych

Sarah E. Graham

sarahgra@umich.edu

ORCID iD: 0000-0003-1271-2489

ACKNOWLEDGEMENTS

This thesis would not be possible without the support of all of the teachers and mentors that I have had throughout my lifetime. I am incredibly grateful to them for encouraging a love of science and math that has allowed me to complete this Ph.D. My undergraduate research advisor, Dr. Andrés Cisneros, was an incredible mentor and spent a great deal of time helping me to learn the fundamentals of computational research. This training has given me a solid base of knowledge to build upon since then. During my graduate research, Dr. Heather Carlson and the rest of the members of the lab provided guidance numerous times. Whether they were helping me to think through a problem or to brainstorm ideas for analysis, they were continuously supportive. I am also grateful to Dr. Betsy Foxman, for allowing me to work in her lab in fulfillment of the TREC certificate. The members of her lab were always willing to train me on techniques and equipment that I was unfamiliar with, and made the additional research experience very rewarding and enjoyable. I am also indebted to all of my committee members, Dr. Heather Carlson, Dr. Charles Brooks, Dr. Kevin Kubarych, and Dr. Tomasz Cierpicki for their insightful feedback on my research progress. I am especially thankful for the computational resources provided by the Brooks lab here at the University of Michigan. Their GPU cluster enabled many of the experiments and results described in this dissertation. I am also appreciative to all of my friends and family who have provided support and friendship during what sometimes feels like a never-ending journey. And most of all, I am grateful to my husband Mark, for supporting me daily and taking care of our family during the early mornings and late nights of graduate school.

TABLE OF CONTENTS

LIST OF TABLES

Table:

LIST OF FIGURES

Figure:

LIST OF APPENDICES

ABSTRACT

Mixed-solvent molecular dynamics (MixMD) is a cosolvent mapping technique for structure-based drug design. MixMD simulations are performed with a solvent mixture of small molecule probes and water, which directly compete for binding to the protein's surface. MixMD has previously been shown to identify active and allosteric sites based on the time-averaged occupancy of the probe molecules over the course of the simulation. Sites with the highest maximal occupancy identified known biologically relevant sites for a wide range of targets. This is consistent with previous experimental work identifying hotspots on protein surfaces based on the occupancy of multiple organic-solvent molecules. However, previous MixMD analysis required extensive manual interpretation to identify and rank sites. MixMD Probeview was introduced to automate this analysis, thereby facilitating the application of MixMD. Implemented as a plugin for the freely available, open-source version of PyMOL, MixMD Probeview successfully identified binding sites for several test systems using three different cosolvent simulation procedures. Following identification of binding sites, the occupancy maps from the MixMD simulations can be converted into pharmacophore models for prospective screening of inhibitors. We have developed a pharmacophore generation procedure to convert MixMD occupancy maps into pharmacophore models. Validation of this procedure on ABL kinase showed good performance. Additionally, we have identified characteristic occupancy levels for non-displaceable water molecules so that these sites may be incorporated into structure-based drug design efforts. Lastly, we have explored the potential for accelerated sampling methods to be used in tandem with MixMD to simultaneously capture conformational changes while mapping favorable interactions within binding sites. These developments greatly extend the utility of MixMD while also simplifying its application.

In addition, two exploratory studies were completed. First, traditional MD simulations were performed to understand the dynamics of NSD1. Crystal structures of NSD1 capture the post-SET loop in an autoinhibitory position. MD simulations allow conformational sampling of this loop, yielding insight into its dynamic behavior in solution. Second, an epidemiological study was conducted which was aimed at understanding the transmission and sequence variation of CTX-M-type β-lactamases, in fulfillment of the clinical research component of the MICHR Translational Research Education Certificate.

Chapter 1. Introduction

1.1 Protein Sequence, Structure, and Function

Protein function requires a delicate balance of sequence and structural motifs with conformational dynamics. Important sequence elements, such as catalytic residues, are generally conserved across protein families as these elements have evolved in concert with the protein's function. However, there may be a great deal of sequence variability even between highly related proteins, which can be important in regulating activity. For example, the kinase class of enzymes is responsible for transferring phosphate groups from an adenosine triphosphate (ATP) cofactor to a specific residue on a target protein. While there are a number of amino acids that are conserved between individual kinases, including the Aspartate-Phenylalanine-Glycine (DFG) motif that coordinates binding of Magnesium ions and ATP, there exists a great deal of sequence and structural variation outside of this region[1]. At first glance these differences may not seem important, as they are not necessarily involved in catalysis, but in fact are critical to allow proteins to function both specifically and efficiently within the context of the cell. There are a vast number of pathways involved in cellular and organismal function, with individual proteins having a specific role within this network. Understanding the role of sequence and structural variations in garnering this specificity is essential to understanding disease processes and in identifying how small molecules could potentially inhibit these functions.

The combination of genetic and epidemiological studies with biochemical research has generated a wealth of information about numerous proteins and cellular pathways, but there are still many unanswered questions. We do not yet know all of the members of these pathways or all of the interactions taking place. There is also heterogeneity within the protein

sequences carried between individuals. These differences may have subtle effects, or may have very detrimental effects, and in some cases may emerge upon treatment with small molecule inhibitors. This is especially relevant within the development of antiviral, antibacterial, or anticancer treatments, where resistance to the inhibitors is a significant factor limiting treatment of these diseases[2, 3]. In the case of antibiotic development, bacteria have evolved a number of resistance mechanisms, including the production of enzymes which chemically alter drug molecules to prevent them from reaching their intended target. The specificity of these enzymes for different classes of inhibitors is dependent on the enzyme's sequence, and may evolve over time to yield increasing levels of resistance. Traditionally, these differences in enzyme structure and function would be examined using a combination of structural and biochemical experiments. For example, NMR or crystallography studies can be used to understand differences in protein structure, while biochemical assays can be used to measure differences in enzyme kinetics. However, performing these experiments to characterize every protein variant is simply not feasible, due to time, financial, and experimental limitations. Alternatively, experimental data can be combined with computational simulations to examine the contribution of individual amino acids to protein dynamics and function, thereby allowing for the study of a much larger number of variants while bypassing the experimental limitations.

1.2 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations are a promising means to understand the role of individual amino acids in regulating protein structure and dynamics. For example, many proteins have conformational changes which are important to their function[4]. Molecular dynamics simulations can be used to simulate these conformational changes in atomic level detail, allowing researchers to study the contribution of individual amino acids to these processes[5]. The results of these simulations can be utilized in many ways, depending on the goal of the study. For instance, the resulting structures can be used to focus biochemical studies, such as mutagenesis, on regions or domains of interest to confirm the role that specific residues have been predicted to play. These simulations may also be used to inform structure-

based drug design efforts, by identifying conformational states that may be selectively targeted by inhibitors, or in combination with other techniques to identify potential binding sites on the protein's surface.

In order to perform a molecular dynamics simulation, structural data on the protein or biological system of interest is required. Typically simulations are initiated from a crystal structure of the protein or system of interest, although models of the protein may be created from related proteins when structural data is unavailable. Files containing descriptors of every atom within the protein are then created, which allow for interactions between protein atoms or between protein atoms and solvent to be calculated based on the bonded (bond lengths, angles and dihedral angles between atoms) and non-bonded interactions (Van der Waals and Coulomb interactions) as determined from the starting structure. The total potential energy is then given as the sum of each of these individual terms, as shown in Equation 1 for the AMBER suite of programs:

$$V(r) = \sum_{bonds} K_b\,(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 \qquad (1)$$

$$+ \sum_{dihedrals} (V_n/2)(1 + \cos[n\varphi - \delta]) + \sum_{non-bonded} (A_{ij}/r_{ij}^{12}) - (B_{ij}/r_{ij}^{6}) + (q_i q_j/r_{ij})$$

Using the relationship between force and potential energy, $F = -\nabla V = ma$, allows for the positions and velocities of the atoms within the system to be calculated over time. In order to solve for the velocity and position at each point in time, the equations of motion must be numerically integrated, as there is no exact solution. The time step that can be used is limited by the fastest motions in the system, such as bonds containing hydrogens. This requires time steps on the order of 1-2 fs, which limits the normally accessible simulation timescales. A number of programs have been developed to carry out these simulations, including AMBER, CHARMM, and NAMD[6-8]. Typically, many individual simulations are done and then average motions may be analyzed over time to understand the specific interactions of interest.

Simulation times usually range from tens to hundreds of nanoseconds, depending on the size and potential conformations of the chosen system. These timescales allow researchers to capture dynamic changes such as side chain rearrangements, loop motions, and helix bending. Traditionally, these simulations are run across clusters of cpus (central processing unit), and/or gpus (graphics processing unit) to achieve simulations of this magnitude within a reasonable time period. For simulations of biological processes that occur at longer timescales, specialized methodologies have been developed. The Anton machine has been developed by D.E. Shaw Research to enable simulations of up to a millisecond, which allows researchers to capture longer timescale processes such as protein folding[9]. For example, Anton was used to perform molecular dynamics simulations of the fast folding protein gpW in tandem with NMR experiments to understand the contributions of interacting residues to the folding process[10]. Distributed computing methods are also notable for being able to tackle long timescale processes by harnessing the power of huge numbers of cpus and gpus. The Pande group has developed the Folding@home project which uses thousands of computers to parallelize molecular dynamics simulations[11]. In the Folding@home project, individuals "volunteer" their computers to perform calculations for the Folding@home team when the computers would otherwise be idle. The group has successfully applied this methodology to several systems, including Src Kinase. In this study, they generated 500 μs of total simulation time from approximately 24,000 individual simulations. The use of Markov state models allowed the group to analyze the extremely large trajectories in order to understand the dynamics and activation of Src, and resulted in the identification of a potentially druggable intermediate state[12].

Enhanced Sampling Techniques

As an alternative to running traditional molecular dynamics simulations on very long timescales, specialized sampling methods can be used to accelerate conformational sampling, or to focus sampling on a particular conformational transition of interest. As shown in **Figure 1.1**, the probability of transitioning between two states at different energy levels depends on

the energy barrier or "barrier height" between the two states. The rate to cross this barrier is given by Equation 2:

$$k \propto e^{-\Delta G/RT} \qquad (2)$$

where ΔG is the change in energy between the starting state and the intermediate state (the barrier height), R is the gas constant, and T is the temperature[13]. At the most basic level theory wise, sampling may be accelerated by increasing the temperature of the simulated system. Temperature-accelerated molecular dynamics was introduced by Sørensen and Voter in 2000[14]. In this method, transitions between states are accelerated due to the increased temperature. As this method can potentially sample high energy states that wouldn't normally be accessible, only transitions that would occur at the desired temperature are kept and others are filtered out[14]. Initial applications of this method focused on simulations of atoms on a solid surface, but temperature-accelerated simulations have since been extended and applied to proteins[15]. Replica-exchange molecular dynamics, or REMD, similarly uses multiple temperatures to simulate the system of interest[16]. In REMD, multiple replicas of the system are simulated at varying temperatures. Replicas at neighboring temperatures may be exchanged, effectively allowing the system to bypass barriers that would exist in traditional molecular dynamics simulations.

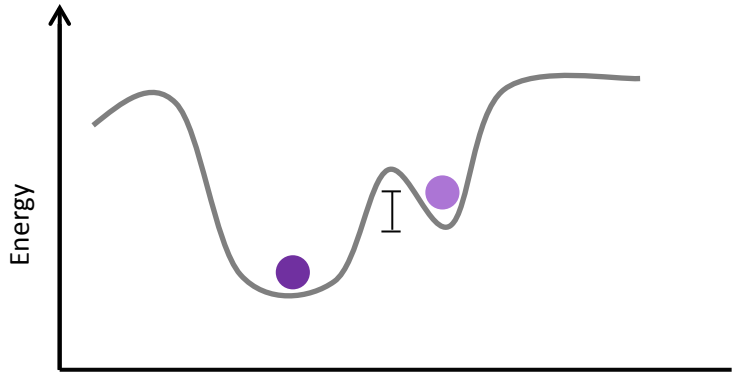


Figure 1.1: The probability of transition from one state (dark purple) to another (light purple) depends on the energy barrier between the two states.

In cases where the region of conformational space to be sampled is known in advance, specialized methods can be used to focus sampling in this area. In the umbrella sampling method, introduced by Torrie and Valleau in 1977, a bias is used to force the system to sample over energy barriers[17]. In this method, the transition of interest is divided into a number of smaller changes, called "windows". For example, if one wishes to study the dihedral angles for a small peptide, the potential conformations may be separated into smaller increments, for instance every 10°, and each small conformational transition may be simulated simultaneously. During the simulations, restraints are used to ensure that the conformational sampling remains near the desired region. The resulting trajectories can then be combined, such as with the weighted histogram analysis method or WHAM, to yield the unbiased energy over the entire reaction coordinate[18].

In cases where the desired end states are not known in advance, molecular dynamics simulations can be accelerated using a modified potential energy surface. If a bias is added to the potential energy surface such that the depths of the wells are decreased, the effective barrier height separating different states is decreased. In the hyperdynamics method, introduced by Voter in 1997, a bias potential is used to increase the potential energy within wells, while leaving the transition state regions at the original potential energy[19]. This allows for the systems to increasingly sample transitions over the barrier regions. In a similar manner, the McCammon group has introduced the accelerated molecules dynamics (aMD) method[20, 21]. In aMD, a potential energy boost (ΔV(r)) is added to the potential energy whenever the system's energy drops below a predetermined cutoff value (E), given in equation 3.

$$V^*(r)\begin{cases} V(r), & V(r) \geq E \\ V(r) + \Delta V(r), & V(r) < E \end{cases} \qquad \Delta V(r) = \frac{(E - V(r))^2}{\propto + (E - V(r))} \qquad (3)$$

As shown in **Figure 1.2**, the potential energy boost decreases the depth of the well, effectively decreasing the barrier height and promoting sampling between nearby states. The level of boost in controlled by the tunable parameter, α. As shown in Equation 3, the potential energy boost is inversely related to the value of α, with smaller levels of α yielding larger levels of

boost. The level of boost can also be controlled by varying the energy cutoff E, although one must be careful not to set the cutoff value so high that the system is essentially exploring a flat energy landscape.

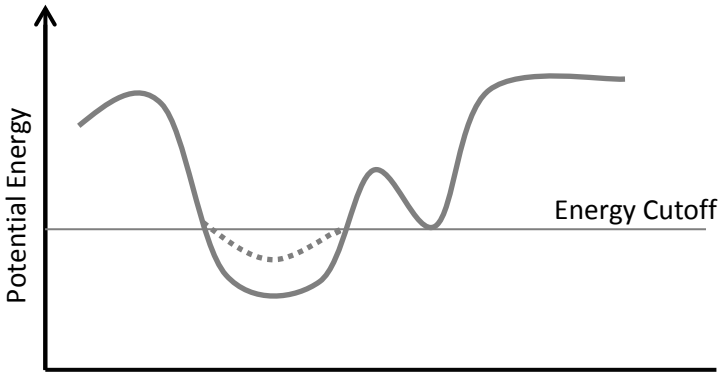


Figure 1.2: In the accelerated molecular dynamics method of McCammon, a boost is added to the potential energy when it is below a specified energy cutoff, which effectively decreases the barrier between related conformations. In regions below the energy cutoff, the system evolves according to the modified, "boosted" potential energy surface, depicted as the gray dashed line.

Accelerated molecular dynamics provides two mechanisms of boosting the sampling during the simulation. The boost may either be applied to the potential energy alone, as in Equation 3, or may be applied to both the dihedral energy and the potential energy[22]. This "dual-boost" method was shown to sample conformational space of an alanine dipeptide more efficiently than either the potential or dihedral energy boosts alone while still maintaining the expected sampling distribution[22]. One caveat with the application of aMD is that the resulting distributions must be reweighted to recapture the original potential energy surface. Reweighting algorithms have been developed that can effectively recapture the underlying potential energy surface for small systems, but are less accurate when applied to large systems[23].

This method has been successfully applied to understand conformational sampling in a number of systems. For example, Guo and Zhou applied accelerated molecular dynamics simulations in combination with conventional MD to study the mechanism of allosteric

communication in the regulatory subunit of protein kinase A (PKA)[24]. PKA has two binding sites for cAMP that sequentially unbind to deactivate the enzyme, but the molecular mechanisms governing this ordered process were previously unknown. Simulations of cAMP bound to either one or both of the binding sites in PKA in comparison with apo simulations of PKA allowed the authors to identify the pathway underlying the allosteric communication and implicated a role for essential interactions with a tryptophan residue that stacks with one of the cAMP molecules[24]. aMD has also been applied to the study of thrombin and its interactions with individual domains of thrombomodulin[25]. Previous studies had suggested the role of thrombomodulin binding in altering conformational loop dynamics in the active site region of thrombin. Residue-residue correlation analysis from the resulting trajectories identified allosteric communication pathways which explained the observed anticoagulant activity differences between binding of thrombomodulin56 and thrombomodulin456 to thrombin.

It is evident from these studies, as well as many others, that molecular dynamics simulations offer great insight into the function and dynamics of biological systems. These simulations also give insight into the ways by which biological function and activity might be modified by small molecule inhibitors. Once the dynamics and function of these enzymes are thoroughly understood, small molecule inhibitors that act to block or enhance this function may be designed. For example, molecular dynamics simulations may give insight into inactive conformations that could potentially be stabilized by ligands to block enzymatic activity, or may help to identify important interactions within an active site region that could preferentially interact with an inhibitor over the natural ligand.

1.3 Computational Drug Design

Since the rise of computational technologies in the 1970's and 1980's, computational methods for drug discovery have had widespread use in both commercial and academic studies towards the design and selection of potential small molecule inhibitors. In the most general sense, computational techniques may be used to understand differences between small

molecules and to rationalize the patterns of activity seen within related series of compounds. QSAR (quantitative structure-activity relationship) methods are frequently used to understand trends in ligand activity seen between similar classes of ligands[26]. QSAR combines both statistical and computational methodologies in order to examine hundreds to thousands, or even millions of compounds for potential activity. For example, some of the earliest applications of QSAR were done by researchers wanting to understand how changing substituents on a common scaffold affected activity[27, 28]. Using such descriptors as compound hydrophobicity (based on octanol-water partition coefficients, logP) and steric effects (calculated from a compound's van der Waals radius), researchers were able to analyze trends in activity and suggest guidelines for the selection of compounds to synthesize. QSAR techniques are also used in the clustering of molecules, to identify closely related compounds and to identify diverse sets of compounds for testing in order to efficiently cover chemical space.

Docking methods that attempt to predict the interactions between an individual ligand and a receptor are also being developed and are widely used[29]. These methods can be applied in multiple ways, depending on the goal of the study. In cases where a ligand is known to be active and a protein structure is available, but 3-D structural information of the protein-ligand complex doesn't exist, docking can be used to predict the binding conformation of the ligand. Docking can also be applied as a virtual screening methodology to rank ligands for their potential to bind to a specific target. The docking procedure consists of two main steps, pose generation followed by scoring and ranking of the generated poses. In order to identify the poses, or orientations of the ligand within the receptor binding pocket, the potential configurations of the ligand must be identified. In some methods, the potential ligand conformations are generated concurrently during the process of ligand placement. Following ligand placement within the receptor binding site, the strength of the interactions for each conformation is assessed in a process known as scoring. Several classes of scoring functions exist, which vary in complexity[29]. Typically faster, less accurate scoring function are used for initial screening of compounds with more complex and computationally demanding scoring

functions reserved for secondary screening. Many docking programs are available, both freely and through commercial licenses, including MOE, Glide from Schrödinger, DOCK, and AutoDock[30-34]. Docking studies are frequently applied as intermediate steps in larger drug discovery efforts, but may sometimes be used as the primary technique. For example, in a recent study by Chiem et al., the Glide docking program was used to virtually screen for inhibitors to AAC(6')-Ib, an enzyme that confers resistance to aminoglycoside antibiotics[35]. The authors initially docked 280,000 compounds from the ChemBridge database into a crystal structure of the enzyme using the standard precision method in Glide, followed by the extra precision method. The authors chose 78 compounds for testing, which yielded one active compound[35].

In a similar way, series of ligands that bind to a desired target may also be examined through the use of ligand-based pharmacophore models to identify common features that are responsible for a ligand's activity[36]. Such methods follow a general workflow where known ligands are selected, potential conformations are enumerated, their structures are aligned, and overlapping common features are identified. In the absence of crystal structures that could indicate the exact interactions that a ligand is making with a receptor, such models are a good way to infer which interactions are crucial to a ligand's activity. The resulting consensus features may be used to rationalize ligand activity and for the screening of potential ligands for synthesis and experimental testing. A number of software packages to perform ligand-based pharmacophore modeling are available, including commercial packages such as MOE and the PHASE utility from Schrödinger[30, 37, 38]. For example, the PHASE program incorporates hydrogen bond donors and acceptors, hydrophobic, aromatic, and positively and negatively charged interactions[37, 38]. Such methodologies are termed ligand-based drug design for their focus on the properties of ligands in mediating receptor-ligand interactions. One of the caveats of using such ligand-based methods is the requirement for known active ligands. Known ligands will not necessarily represent all of the interactions that a receptor is capable of making, and may therefore limit the potential discovery of new compounds.

Alternatively, receptor-based pharmacophore models may be used. Rather than examining consensus features solely between known ligands, receptor based methods look for potential interactions that the receptor may make, in order to identify all potential interactions that could be made with a hypothetical ligand. The resulting pharmacophore models can then be applied in a very similar manner to those generated using ligand-based methods, but the process by which they are generated differs. For example, Waltenberger et al. used an iterative procedure with LigandScout and Discovery Studio from Accelrys to generate both ligand-based and structure-based pharmacophore models of soluble epoxide hydrolase[39, 40]. Starting with 9 crystal structures and 68 active compounds identified from the literature, they were able to develop several pharmacophore hypotheses. The resulting models were screened against the Specs database to identify potentially active compounds. The compounds were ranked by their degree of fit to the pharmacophore model, and inspected for potential steric clashes with the receptor and the presence of non-desirable functional groups which led to the selection of 48 compounds for experimental testing. Of the tested compounds, 19 were active inhibitors[40]. The identification of active compounds relative to inactive compounds when screening ligands is termed the "hit-rate". In traditional high-throughput screening approaches, the hit rates are universally much lower than that observed in this study, showing the promise and utility of these structure-based drug design techniques. Several programs exist that are able to directly create pharmacophore models using either receptor or receptor-ligand complexes, including Schrödinger, and the previously mentioned LigandScout and Accelrys [39, 41, 42]. In addition to these "all-in-one" programs, a number of complementary methods have been developed to identify potential interactions that a receptor may make. These methods, which will be discussed in greater detail in the following section, do not generate pharmacophore models directly, but rather focus on identifying all of the potential interactions that a receptor may make.

1.4 Mapping Binding Sites on Protein Surfaces

It has been shown that binding sites on a protein's surface can be identified and characterized through their interactions with small molecules. For example, the multiple solvent crystal structure (MSCS) technique uses crystallography in combination with various solvents, such as acetonitrile, ethanol, and isopropyl alcohol, to look for potential binding sites[43, 44]. MSCS was first applied to the protein elastase with acetonitrile as the solvent[45]. Later studies by the same group solved structures of elastase with additional solvents[43]. As shown in **Figure 1.3**, some of the probe molecules bind in single sites, but many of them overlap within specific regions on the proteins surface. Initial studies of this method showed that these molecules preferentially cluster within binding sites on the protein's surface, indicating regions of the protein which favorably interact with multiple functional groups[45]. Such "hotspots", i.e. regions that bind multiple probe molecules, are therefore indicative of binding sites and may be used to find easily desolvated regions which can potentially bind small molecule inhibitors. Indeed, ligand bound structures of elastase show inhibitors binding at this site.

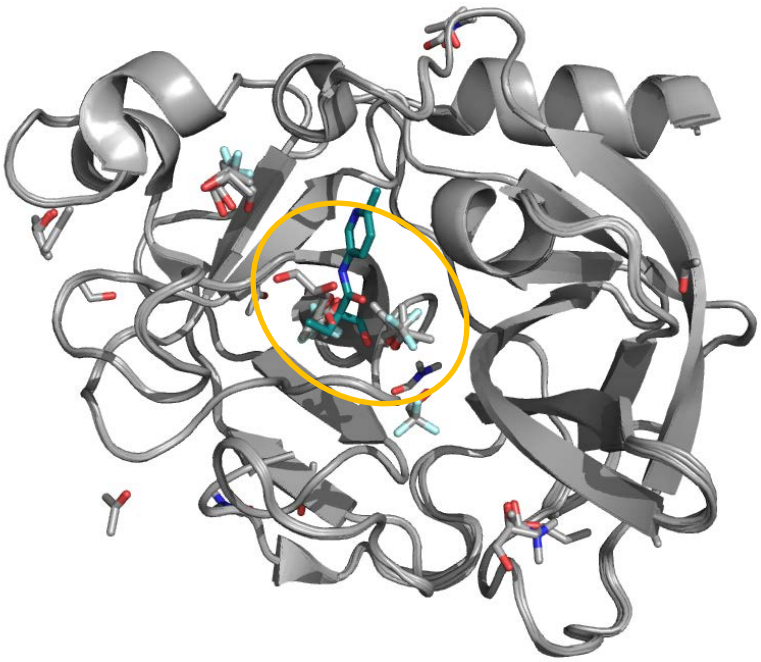


Figure 1.3: Multiple solvent crystal structures of elastase are shown. It can be seen that regions on the protein's surface that have multiple overlapping probe molecules correspond to inhibitor binding sites. Probe molecules are shown in gray, taken from nine different MSCS with the inhibitor JM102, (PDB: 4YM9, unpublished) shown for reference[43].

NMR methods have similarly been developed to map binding sites of small molecule probes. For example, in the SAR (structure-activity relationship) by NMR approach, compounds are screened via NMR for binding to a protein of interest[46]. The first SAR by NMR study screened libraries of compounds for binding to FK506 binding protein (FKBP). In the initial study, 10,000 compounds dissolved in perdeuterated DMSO were screened. Compounds that bind to the ^{15}N labeled protein induce a ^{15}N or ^{1}H-amide chemical shift change in the 2-D ^{15}N–HSQC spectra relative to the apo protein, and can therefore be identified. Once an initial ligand is found that binds to the protein (a "hit"), derivatives of the initial ligand are screened and optimized in an iterative procedure that is repeated until a "lead" compound is found that has high affinity. In the case of FKBP, this process was repeated to identify high-affinity binders for a nearby site. These two high-affinity compounds which bind in adjacent regions on the protein surface were then linked, yielding a larger high-affinity (nanomolar) compound. One of the advantages of this method is the ability to simultaneously determine the binding affinity of the molecules, so that compounds may be compared to identify the highest affinity binders.

Computational techniques have also been developed to identify favorable binding sites. The GRID method, introduced in 1985 by Goodford, calculates the potential energy of a probe molecule interacting with a protein by considering the interactions at each xyz point on a 0.5 Å gridded representation of the protein surface[47]. Favorable binding sites on the protein surface can then be determined by examining the predicted affinity at each grid point location. As a test case, the GRID method was applied to several test systems, including *E. coli* dihydrofolate reductase (DHFR). Regions with favorable interaction energy on the surface of DHFR determined with GRID correspond to the known binding site of trimethoprim. Similar computational techniques have been developed which make use of multiple probe molecules. In the multiple-copy simultaneous search method (MCSS), many small molecule probes are distributed across a protein surface and minimized to identify the most favorable binding sites[48]. The method was first applied to influenza hemagglutinin. Acetate, methanol, methane, methyl ammonium, and water were selected as representative functional groups and 1,000 to 5,000 copies of each probe were randomly distributed within the binding site of hemagglutinin.

Minimization of the probes yielded ~25-125 minima, depending on the probe type, and showed good agreement with the known sialic acid binding site.

More recently, the FTMap computational method was introduced, which uses sixteen different probe types to identify binding hotspots on the protein surface[49]. FTMap is utilized as a webserver where users upload their protein, DNA, or RNA target. The receptor is then flooded with probe molecules, currently ethanol, isopropanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide and N,N-dimethylformamide. The probe molecules sample different orientations through rotational and translational motion followed by minimization to identify the most energetically favorable binding sites. The locations of the probe molecules are clustered to identify discrete binding sites on the receptor's surface. The resulting sites are then ranked by average energy and number of clusters that they contain to identify the top ranked site, as well as smaller secondary sites. One of the main benefits of FTMap is its speed and ease of use, allowing for hot spot predictions within a manner of hours or days. In one recent study of CXCL12, FTMap was used to identify the main hotspots mediating protein-protein interactions between the chemokine CXCL12 and the CXCR4 peptide[50]. FTMap identified several sites that corresponded to the binding sites of specific CXCR4 residues. Following the initial insight from the FTMap calculations, the authors utilized docking to screen the ZINC database for potential inhibitors capable of blocking the CXCL12-CXCR4 interaction. Experimental testing confirmed the ability of these ligands to bind to CXCL12 at the three main hot spots identified by FTMap, demonstrating the utility of this method.

While these methods are very computationally efficient, they tend to neglect important determinants of ligand binding, namely the role of solvent and protein flexibility. For example, it has been shown that FTMap is able to identify active sites, but may be unable to identify allosteric sites, likely because of the inability to fully capture the conformational states of the protein using static crystal structures as input[51]. The use of static structures in computational drug design has been shown to result in the identification of false minima, therefore making it

difficult to identify the true binding sites on a protein's surface[52, 53]. Several methods have been used to incorporate protein flexibility into structure based drug design, including the use of ensembles of crystal structures or the use of side chain rotamers[54]. For example, within the FTMap suite, FTFlex allows for side chain flexibility and FTDyn allows for ensembles of crystal structures[49]. However, these methods are not always sufficient to capture the full conformational flexibility of the proteins. For example, crystal structures may not be available of all potential conformations, such as in cases where apo crystal structures (those containing no ligands) are unavailable. There may be conformational rearrangements that occur upon ligand binding between the ligand-bound and apo protein structures which are not fully represented by available ligand bound structures. Likewise, the role of competition between water and probes is typically neglected in these methods. Prior to ligand binding, proteins are surrounded by solvent molecules. In order for a ligand to bind, these water molecules must be displaced, or may be conserved and bridge interactions between the protein and ligand[55-60]. It has been shown that structure-based drug design methods that explicitly consider the role of water molecules have better prediction rates than methods which neglect water molecules[61]. In order to overcome these limitations, molecular dynamics simulations incorporating multiple solvent probe molecules have been developed, which are able to explicitly account for protein flexibility and the role of solvation in mediating protein-ligand interactions.

1.5 Cosolvent Molecular Dynamics Simulations

Cosolvent simulations have been developed as a promising means to capture protein flexibility while simultaneously accounting for the role of water. In these simulations, the protein is immersed in a box containing water and small molecule probes, and then subjected to molecular dynamics simulations. The small molecule probe types differ between methods, but are typically chosen to represent common functional groups found in drug-like molecules. Several methods have been developed, which differ in their experimental setup and analysis. The most relevant and well developed methods will be discussed in detail below; a comprehensive review of all cosolvent simulation methods is available[62].

SILCS

The site-identification by ligand competitive saturation (SILCS) method by the MacKerell group is particularly notable, as the method has been extended for the development and screening of novel ligands. The first SILCS study was published in 2009, and focused on the mapping of the SMRT and BCOR peptide interaction sites on the BCL-6 oncoprotein[63]. In the initial SILCS methodology, the protein was solvated in a box containing propane and benzene molecules, each in a ~1M concentration with water. Propane and benzene were chosen to identify hydrophobic and aromatic interactions, respectively, and water molecules identified potential hydrogen bonding interactions. Ten simulations of 5ns production time each were completed, to yield 50ns of total simulation time. In order to prevent aggregation of the solvents, an artificial repulsion term was applied when two fragments came within a set distance. In the SILCS methodology, this is implemented by modifying the Lennard-Jones interactions to include an additional term for the center of each benzene and propane, such that the probe molecules will repel each other at short distances. The resulting trajectories were analyzed at 2ps intervals to yield the occupancy of the probe molecules over the simulation at every point on a 1 Å cubic grid. The resulting occupancy maps showed good agreement with many of the known residues in SMRT and BCOR that are known to be important determinants of binding[63]. One of the main drawbacks of the initial SILCS methodology was the limited number of probe types used. For example, using water as the probe for hydrogen bond donor and acceptor interactions prevents one from distinguishing which water interactions may be preferentially displaced by other functional groups capable of making similar interactions.

The next subsequent updates to the SILCS methodology focused on the development of descriptors for the binding affinity of probes. Using the Boltzmann relationship:

$$\frac{k_i}{k_j} = e^{-\beta(E_i - E_j)} \qquad (4)$$

where k_i and k_j are the number of molecules in states i and j, respectively, E_i and E_j are the associated energies, β is equal to $1/k_BT$ where k_B is Boltzmann's constant and T is temperature, allows for the calculation of energetic differences between states based on the occupancy of the states. In the SILCS methodology[64], this is calculated on the basis of individual grid points:

$$GFE^{T}_{xyz} = \min\left\{-RT \ln \frac{occupancy^{T}_{x,y,z}}{\langle bulk\ occupancy \rangle}, 0\right\} \quad (5)$$

where the GFE, or grid based free energy for a specific grid point (x,y,z) and specific functional group/probe type T, is calculated from the occupancy of the probe type at that grid point relative to the occupancy in bulk solvent. In the initial SILCS application of this equation, if the occupancy based free energy was greater than zero, meaning the occupancy at some (x,y,z) point was less than the expected bulk occupancy, the GFE was assigned to be zero[64]. This equation was then used to predict the binding affinity of a ligand (Ligand Grid Free Energy, of LGFE) by summing over the GFE scores for each atom type in a ligand calculated, as shown in **Figure 1.4**. To test this methodology, Raman et al. performed simulations of trypsin, α-thrombin, HIV protease, FKBP12, Factor Xa, NadD, and ribonuclease A in the presence of a 1 M propane and benzene solvent mixtures[64]. In the present method, 200 ns of production simulation were completed for each system. The ability of the method to correctly rank the crystallographically observed binding mode was then tested by comparing the predicted LGFE of the crystal binding form of the ligand to decoy positions. Overall, the method tended to predict the crystallographically observed binding mode of the ligands as energetically favorable, but not necessarily as the top ranked pose.

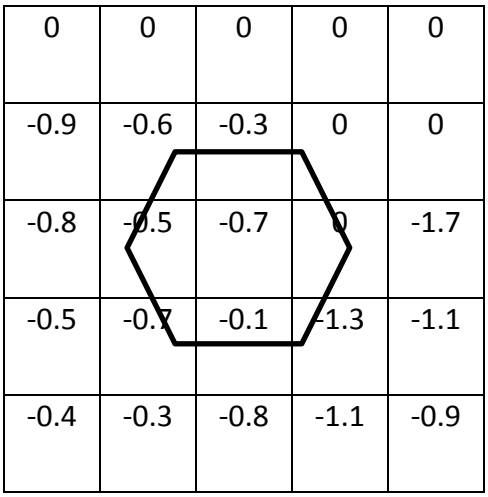


Figure 1.4: The ligand grid free energy in the SILCS methodology is calculated by considering the positions of each of the atoms in a ligand of a specific type and the associated grid free energy values for each position (kcal/mol). The grid free energy values are then summed, to give the overall ligand grid free energy value[64].

In order to increase the efficiency of solvent sampling within the SILCS methodology, a combined grand canonical-like Monte-Carlo/molecular dynamics (GCMC-MD) approach was introduced[65]. In the SILCS GCMC-MD procedure, solvent probe molecules and water are inserted into the system from a reservoir for some number of GCMC steps, followed by molecular dynamics simulations to allow sampling of the solvent molecules with the protein. This cycle is repeated several times to allow for efficient sampling of the solvent molecules with the protein. Initial validation studies of the SILCS GCMC-MD method focused on the occluded binding pocket of T4 lysozyme, with subsequent studies applying the methodology to successfully identify binding sites in androgen receptor, peroxisome proliferator-activated receptor gamma, metabotropic glutamate receptor, and beta-2 adrenergic receptor[65, 66].

In 2013, Raman et al. published an update to the SILCS methodology which addressed the deficiency of limited probe types, and extended the SILCS LGFE metrics to consider multiple conformations of a ligand[67]. Methanol, formamide, acetaldehyde, methylammonium, and acetate were added to the original benzene and propane probes to form the SILCS tier II probe set. To validate the additional probe molecules, their group selected Factor Xa, P38 Map Kinase, RNase A and HIV protease as test systems. In the SILCS tier II method, the seven probes

are simulated together, at approximately 0.25 M each. For each system, 10 simulations of 40ns were performed. Following simulation, the trajectories are aligned and the occupancy of the probe molecules is determined as previously described. In order to consider conformational sampling of the ligand, the authors included a Monte Carlo sampling step for the ligands within the resulting occupancy maps. The LGFE is then calculated in the same way as previously described for each conformation, with the exception that the maximum GFE value was assigned to be 3 rather than 0, and that weighted LGFE values are calculated using a coefficient that varied depending on the probe type. The authors rationalize the coefficients as accounting for the additive effect of including multiple atoms in the LGFE calculations, such that benzene LGFE values would be divided by 6 to account for the 6 non-hydrogen atoms in a benzene molecule. The LGFE values for each conformation k are then combined according to the following equation:

$$LGFE = -k_B T \ln \langle exp\left(-\frac{LGFE(k)}{k_B T}\right)\rangle \qquad (6)$$

Depending on the input ligand conformations used, R^2 values for the correlation between LGFE values and experimental binding affinities ranged from 0.02 to 0.79. While the method had good performance in predicting binding affinities for some targets, others, such as Factor Xa (R^2 range 0.02-0.33), were not predicted well by the present methodology. Most recently, the SILCS methodology has been extended to compute relative binding affinities between related compounds, using either a single-step free energy perturbation (SSFEP) method or through a comparison of LGFE values[68, 69]. A comparison of FEP, SSFEP, and the SILCS LGFE was performed on ACK1 and MAP kinase, and showed that SSFEP and SILCS LGFE gave comparable results with the more time-consuming FEP method, highlighting the potential of cosolvent simulations in facilitating fragment based drug discovery efforts by guiding the selection of suitable molecules for synthesis and testing.

Notably, the SILCS method has also been extended to develop pharmacophore models that can be used for virtual screening[70, 71]. In the SILCS pharmacophore methodology, a grid

based occupancy is used as previously described, in which each grid point is given a grid free energy (GFE) value as shown in Equation 5. The probe maps are then visualized using varying GFE cutoffs, to yield favorably occupied regions for each probe type. Based on manual selection, regions are chosen for inclusion into the pharmacophore model. Highly occupied sites within the region of the protein chosen for pharmacophore modeling are then converted into pharmacophore features. This is done by clustering the grid points into distinct sites, with the center of the cluster becoming the center of the pharmacophore feature, and the radius of the pharmacophore feature set to include all grid points within the cluster. In cases where an aromatic feature interacts with a hydrophobic feature by more than half, the features are combined to yield an aromatic-hydrophobic feature with the new center being the center of all grid points within the two features and the radius set to encompass all points within the two features. The possible combinations of all pharmacophore features into a pharmacophore model are then compared, based on the sum of the GFE values for the individual pharmacophore features contained in the model. The pharmacophore model can then be used for virtual screening. The most recent version of the SILCS pharmacophore method applied to 8 proteins showed equivalent or better performance in virtually screening potential ligands relative to DOCK 4.0, AutoDock 4, and AutoDock Vina[70].

MDmix

Barril and coworkers have developed the related MDmix method of cosolvent simulations. In the first development of their method, published in 2009, cosolvent simulations of isopropyl alcohol and water were used to identify potentially druggable sites on a protein's surface[72]. In their method, pre-equilibrated boxes of water and isopropyl alcohol at a concentration of 20% v/v are used to solvate the protein of interest. Molecular dynamics simulations of at least 16ns are carried out for each protein, and then the resulting snapshots are overlaid with a grid in a similar manner to the SILCS protocol. The resulting occupancies can then be counted and compared to the expected occupancy, to yield the theoretical binding affinity of the probe molecule for a specific site through the Boltzmann relationship. They estimated the maximal affinity of a ligand for an identified site by clustering highly occupied

grid points and summing the predicted binding affinities within the region. The MDmix method was initially applied to thermolysin, p53, elastase, MDM2, LSA1, protein tyrosine phosphatase 1-b (PTP1B), P38 map kinase, and androgen receptor (AR). The MDmix method correctly predicted the majority of known binding sites, with the exception of the phosphorylated-tyrosine binding site on PTP1B. While the method had good success in locating known binding sites, it suffers from some methodological limitations. A later study by our group using the parameters for isopropyl alcohol from this initial MDmix study showed phase separation between isopropyl alcohol and water, which should not be seen if the parameters chosen are correct[73]. Two additional studies using MDmix have been published, which focus on predicting water displacement and the role of conformational flexibility is the convergence of cosolvent simulations. These contributions will be discussed in subsequent sections.

1.6 The Mixed-Solvent Molecular Dynamics Method (MixMD)

The cosolvent simulation method that has been developed in our group is termed mixed-solvent molecular dynamics (MixMD). Initial studies in the Carlson lab focused on methods to incorporated protein flexibility in computational drug design, which motivated the development of the multiple protein structure (MPS) method[74-76]. In the initial MPS methods, ensembles of protein conformations, either from crystal structures or taken from MD simulations, were used to incorporated protein flexibility in pharmacophore modeling. The initial studies used minimization of probes to map favorable interaction sites on the protein's surface, allowing for the development of pharmacophore models. As a natural extension of this method, the MixMD method has been developed to use molecular dynamics simulations for conformational sampling of the protein while simultaneously mapping favorable interaction sites with cosolvent probe molecules. In 2011, Lexa and Carlson demonstrated the importance of including protein flexibility to accurately map binding sites on protein surfaces[51]. Using hen egg-white lysozyme (HEWL) as a test system, 5 simulations of 10ns with HEWL solvated in a 50% w/w acetonitrile and water mix were completed. HEWL is a particularly good test system, as a crystal structure of HEWL solvated in acetonitrile and water is available, which allows for

experimental validation of the results[77]. Three different restraint schemes were tested to determine the impact of protein flexibility on mapping results; 1) no restraints, 2) rigid restraints on the backbone, and 3) rigid restraints on every atom. Interestingly, the study demonstrated that only the simulations without restraints successully identified the known acetonitrile binding site of HEWL without also identifying spurious sites[51]. Alvarez-Garcia and Barril later studied the effect of using constraints on the observed mapping and binding afinities calculated from cosolvent simulations[53]. Using HEWL in a 20%/80% methanol, isopropyl alcohol, acetonitrile, or ethanol to water ratio, they tested constraints of 1, 0.1, and 0.01 kcal/mol-Å^2 relative to simulations performed without restraints. Increasing levels of constraints on the heavy atoms resulted in more negative ΔG values relative to the unrestrained simulations. The authors propose that weak restraints on heavy atoms will facilitate convergence of the simulations, as the ΔG values calculated from the 0.01 kcal/mol-Å^2 constraints were within error of the values calculated from unconstrained simulations. However, HEWL is known to be a very stable protein, so the generalizabilty of this observed effect is unknown.

The next developments to the MixMD method focused on the development of proper protocols for simulation setup, probe parameters, and analysis[73, 78]. These studies were performed with the goal of identifying procedures that would allow consistent application of the MixMD methodology to proteins with potentially unknown binding sites. Lexa and Carlson initially focused on the study of systems with published multiple-solvent crystal structures[78]. They selected elastase, HEWL, p53 core, RNase A, and thermolysin, all of which had crystal structures containing bound isopropyl alcohol. 5 to 10 50ns simulations of each protein were completed in a 50% w/w acetonitrile or isopropyl alcohol and water mix. Comparing results taken from the first, middle, or end of the trajectories indicated that the first part of the trajectories tends to identify spurious sites. Allowing time for sufficient sampling, and thus analyzing the end of the 50 ns trajectory correctly identified known probe binding sites. Furthermore, the experimentally known probe binding sites were mapped at very high occupancy, of at least 5 standard deviations above the mean occupancy with the 50%/50% w/w

probe to water ratio. Importantly, Lexa and Carlson noted that some of the probe binding sites seen in crystal structures are due to interactions involved in the crystalline environment, and therefore should not necessarily be reproduced by simulations of inidividual proteins. This finding highlighted the need to examine crystal packing contacts and crystallographic density of ligand-bound crystal structures when performing the analysis of cosolvent simulation results.

Lexa, Goh, and Carlson next focused on the development of experimentally consistent probe parameters[73]. As noted previously, the initial MDmix studies by Seco and Barril exhibited unrealistic solvent behavior, specifically the separation of isopropyl alcohol and water into two layers[72]. Lexa et al. tested the behavior of 11 solvents to identify those solvents that were water miscible within the context of an MD simulation with the TIP3P water model[73]. Boxes of 50%/50% w/w mixtures of water and isopropyl alcohol, acetonitrile, acetone, N-methylacetamide, imidazole, pyridine, pyridazine, pyrimidine, pyrazine, benzene, or phenol were examined. Parameters for the probes were taken from AMBER (for acetone, acetonitrile, and N-methylacetamide) or OPLS (all others), as described in the manuscript. In order to quantitatively examine the behavior of the solvent during the simulation, radial distribution functions were calculated between the two solvent types. Radial distribution functions, or RDFs, examine the probability of finding two molecules within a specified distance. At long distances, the RDF should converge to a value of 1, indicating that there is no correlation in the location of the two types, and thus that the two solvent types are sufficiently mixing over the course of the simulation. Using the RDF metric as indicative of proper solvent mixing, Lexa et al. identified isopropyl alcohol, acetonitrile, acetone, N-methylacetamide, imidazole, and pyrimidine as acceptable probe choices that showed water miscible behavior. Notably, benzene, which is used as a solvent probe in the SILCS method, was found to be immiscible with water.

Following meticulous development of the MixMD method to ensure its experimental validity, studies by our group next turned to the application of MixMD in mapping protein surfaces. Ung et al. applied the MixMD method to HIV-1 protease[79]. Structures of both the

semi-open and closed conformations of HIV protease were used. Simulations were performed using acetonitrile-, isopropyl alcohol-, or pyrimidine-water mixtures, at both 50%/50% w/w and 5%/95% v/v ratios. Pre-equilibrated solvent boxes were used for the 50% w/w solvent mixture, while simulations of the 5%/95% mixture used a layered setup. In the layered solvent procedure, the system is first solvated with a layer of the desired probes, followed by a layer of water to yield the desired concentration. The layered procedure was chosen in order to facilitate system setup, while also allowing for competition between probes and water for binding to the protein's surface. Since the 5%/95% ratio of probes to water results in a much lower number of probe molecules relative to the 50%/50% ratio, pre-equilibrated boxes would have necessitated longer simulation times for a sufficient number of probe molecules to sample the protein surface. In the layered procedure, the probes are closer to the protein's surface, facilitating sampling, while the large numbers of water molecules are expected to effectively compete with the probes for binding to the protein's surface, thus yielding realistic results during the simulation. It is important to note that the correct sites were still mapped with either setup procedure. Five simulations of 20 ns were completed for each system, and the last 10 ns were used for further analysis. The resulting trajectories were aligned and the occupancy of probe molecules was determined for each point on the protein's surface using a 0.5 Å cubic grid. The resulting occupancies were normalized by subtracting the mean of the raw data and dividing by the standard deviation, to yield the occupancies in units of 1 standard deviation (equivalent to a Z-score). Both simulation procedures identified known binding sites on the surface of HIV protease, including the eye, face, and exo sites in the open form, and clear mapping of the active site in the closed form. The 5%/95% simulations, however, showed a much greater range between the occupancy of known binding sites compared with the occupancy of spurious minima which facilitated identification of true binding sites. This difference is especially important in the prospective application of cosolvent simulations, as binding sites may not be previously known. The lower probe to water ratio is also more consistent with experimental studies, which are typically performed at much lower concentrations than 50%[79].

Most recently, MixMD was applied to identify allosteric binding sites. Previously, cosolvent simulation methodologies have focused on the identification of the main binding sites of a protein. However, targeting these sites with small molecule inhibitors may not always be feasible, or desirable, depending on the target. In order to circumvent these limitations, inhibitors may target allosteric sites, which may potentially be more amenable to inhibitor development or may provide a greater level of specificity between related targets[80]. Using a layered cosolvent approach at 5%/95% v/v ratio of probe to water, Ghanakota and Carlson applied MixMD to ABL kinase, androgen receptor, CHK1 kinase, glucokinase, PDK1 kinase, farnesyl pyrophosphate synthase, and protein tyrosine phosphatase 1B[51]. All of the starting structures were taken from crystal structures without ligands bound in the allosteric site, to provide the most unbiased test of the MixMD procedure as none of the allosteric sites would be prearranged for ligand binding. Ten independent simulations of 20 ns were completed for each protein, in the presence of acetonitrile, isopropyl alcohol, or pyrimidine and water. As some of the systems bind charged ligands, two additional probe types were introduced: acetate and methylammonium, to identify these interactions. The last 5 ns of each simulation was analyzed to determine the occupancy of each probe on a 0.5 Å cubic grid and normalized as previously described into standard deviation units. Across all systems tested, the allosteric sites were consistently identified as being highly occupied by more than one probe type. After analyzing all systems, Ghanakota and Carlson observed that the top four sites by probe occupancy ranking consistently identified both the active and allosteric sites. Lower occupancy sites typically identified crystal packing interfaces and locations of buffer molecules. For example, in ABL kinase, the myristate binding pocket is the highest ranked site by probe occupancy, followed by mapping of the active site, and later by mapping of the interface of the SH2-kinase domain[51]. This consistent ranking suggests the ability of MixMD to be applied in a prospective manner, to identify other potential allosteric sites as well as active sites.

1.7 Water Prediction Methods

Given the ability of cosolvent simulations to correctly predict probe binding sites, it is likely that such simulations will also correctly predict the conservation or displacement of water molecules upon ligand binding. This has been a long-standing problem in computational drug discovery, as the inclusion of essential water molecules improves predictions[61, 81], but it is not always clear when water molecules should be included and when they may be neglected as likely being displaced upon ligand binding. Initial methods that attempted to predict water displacement were based on statistical analysis of protein-ligand complexes, using various descriptors. In more recent years, methods based on molecular dynamics or Monte Carlo simulations and using more in-depth energetic analysis have emerged.

For example, the Consolv method was one of the first such methods to be introduced that attempted to predict water conservation or displacement upon ligand binding using a knowledge-based approach[82]. Using 13 unrelated proteins that had both apo and ligand bound structures, Raymer et al. selected 157 active-site waters found in the apo structures to act as the test set for Consolv, and 1700 first solvation shell waters (i.e. those directly interacting with the protein, but not necessarily in the active site) for the training set. The waters were then classified as conserved or displaced, depending on their presence in the corresponding ligand-bound structures. The environment surrounding each water molecule was determined, as characterized by the number of nearby protein atoms, the character of surrounding amino acids, the number of hydrogen bonds made between the water molecule and protein, and the crystallographic B-factor. These descriptors were then compared to test water molecules, to identify water molecules having similar characteristics to the water of interest, and to classify the water molecule as conserved or displaced based on this similarity. When applied to the test set, this yielded 77% accuracy in predicting whether a water molecule was displaced or conserved upon ligand binding. Using similar descriptors, García-Sosa and coworkers introduced the WaterScore method, which examined B-factors, solvent-contact surface area, the hydrogen bond energy, and the number of protein-water contacts to differentiate between bound and displaceable water molecules[83]. Using a regression model trained on 14 systems, the WaterScore method predicted the conservation or displacement of water molecules in 4

different test systems with 67.4% accuracy. While these methods are very fast, they may not be able to accurately predict water molecules in all cases, and do not give detailed insight into the interactions that a specific water molecule is involved in.

Alternatively, a number of methods for examining the role and potential displacement of an individual water molecule based on its hypothetical energy have been developed. In order for a ligand to favorably displace a water molecule, the interactions that the displaced water molecule was making must be compensated for. This has led to the development of methods that focus primarily on determining the detailed thermodynamic values for an individual water molecule. For example, Barillari et al. compared the binding affinities of conserved and displaced water molecules using replica exchange thermodynamic integration, and found that displaced waters were bound less tightly than waters which are conserved upon ligand binding[84]. They proposed that the likelihood of a water molecule being displaced or conserved could be calculated by comparing its binding affinity to that of known displaced or conserved waters. Amadasi et al. introduced the HINT/RANK method for predicting water displacement[85]. Using the HINT forcefield and their proposed RANK method, the authors predicted the conservation or displacement of a water molecule based on the number and strength of hydrogen bonding interactions made. When applied to 50 water molecules from 4 proteins, 76% were correctly predicted as displaced or conserved by the HINT/RANK method. The AcquaAlta method was developed to predict the positions of water molecules bridging interactions between proteins and ligands in docking[86]. In this method, waters are placed to optimize hydrogen bonding between the protein and ligand via a bridging water molecule based on distance, interaction energy, and orientational constraints. The AcquaAlta procedure correctly predicted the locations of 76% of waters in the training set, but when applied to docked ligands predicted only 53% of water positions.

Monte Carlo simulations may also be used to analyze the energetics of water molecules. In the Just Add Water Molecules (JAWS) method, Monte Carlo simulations are used to calculate the binding affinity of water molecules[87, 88]. In this method, water molecules are transferred

from bulk solvent to a binding site on the protein's surface. These waters are termed θ waters, where θ = 0 corresponds to a water molecule in the bulk solvent, and θ = 1 is a water molecule at a particular location in the binding site. The probability that a water molecule occupies the binding site rather than the bulk solvent is controlled by the energy difference between the two states. This relationship therefore allows for the binding affinity of a water site to be calculated:

$$\Delta G = -k_B T ln(\frac{P(\theta_i=1)}{P(\theta_i=0)}) \quad (7)$$

based on the difference in sampling for a given site between bound (θ = 1) and bulk (θ = 0) water. This method was successfully applied to a series of ligands for scytalone dehydratase, p38-αMAP kinase, and EGFR kinase to explain differences in ligand binding affinity on the basis of energetics of individual water molecules[88].

Commercial software to predict water binding sites based on an individual water molecule's interaction energy are also available. OpenEye offers the SZMAP method to study the role of individual water molecules[89] In the SZMAP method, an explicit water molecule is placed into a protein-ligand binding cavity, while the rest of the cavity is modeled using the Poisson-Boltzmann implicit solvent model. The water is allowed to sample conformational space over a cubic grid, or may be altered to sample a desired position for comparison to a known potential water site. The energy at each sampled point is then calculated, based on the sum of the van der Waals and electrostatic interactions between the protein, ligand, and water molecules, and a desolvation term for the protein and water. Comparing the energy at a selected point with all other points via a partition function yields the probability of a water molecule being found at an individual point, as given in Equation 8.

$$Probability = \frac{e^{-\frac{E_j - E_{min}}{k_B T}}}{\sum_{j=1}^{N_{orient}} e^{-\frac{E_j - E_{min}}{k_B T}}} \quad (8)$$

The developers of SZMAP also propose a means to calculate the enthalpic and entropic contributions for an individual water molecule's binding affinity, by considering the number of favorable points for a water to occupy, and the strength of the interactions, as given by the numerator of Equation 8[89]. When applied to HIV protease, neuraminidase, trypsin, factor Xa, scytalone dehydratase, and oppA as test systems, SZMAP ΔΔG values calculated for the theoretical conversion of a neutral probe to a water molecule showed good correlation with the replica exchange thermodynamic integration calculated ΔG values. Furthermore, they found the SZMAP calculated entropy of a water molecule to be the best predictor of conservation, with 93% accuracy in the test set.

The SPAM method, developed by GlaxoSmithKline, similarly considers the interaction energy of a water site to calculate a theoretical binding affinity[90]. In the SPAM method, a molecular dynamics simulation is performed to allow for sampling of potential water positions around the protein. Following this, the binding affinity is calculating by summing over the interaction energies of the water molecules during the simulations:

$$Q_{SPAM} = \sum_{E_{water}} [P(E_{water}) \exp\left(-\frac{E_{water}}{RT}\right)]$$

$$G_{SPAM} = -RT \ln Q_{SPAM} \qquad (9)$$

$$\Delta G_{SPAM} = G_{SPAM,bound} - G_{SPAM,bulk}$$

and comparing the resulting to those of bulk water. The SPAM method was applied to study the water in HIV protease that bridges interactions between the "flaps" of the protein and the active-site ligand. It was found that this water had a favorable interaction energy relative to bulk water, but was entropically unfavorable due to its constrained location, and therefore had a net unfavorable $\Delta G_{SPAM.}$ This finding is consistent with the ability of ligands to displace this water molecule and bind with higher affinity[90].

The WaterMap method has been developed by Schrödinger, as an extension of the inhomogenous fluid solvation theory (IFST/IST) methods of Lazaridis[91-95]. IFST based methods focus on the orientational correlation between a solvent molecule and the protein. In the WaterMap method, conformational sampling of the water molecules is achieved via molecular dynamics simulations. The entropy of a water molecule interacting with the protein is then determined by considering the correlations of the water molecule. In combination with the interaction energy, the IFST-based entropy values yield a detailed thermodynamic view of an individual water molecule. This can then be applied to understand potential affinity differences when comparing ligands. For instance, WaterMap was applied to ligands of factor Xa that are known to displace water molecules. WaterMap calculated $\Delta\Delta G_{bind}$ values correlated very well with experimentally known $\Delta\Delta G$ values, with an R^2 of 0.81[94].

IFST based methods have also been further developed academically. Nguyen, Young, and Gilson have introduced a grid based implementation of the IFST method, termed GIST[96]. In order to simplify evaluation of the original IFST equations, the GIST method discretizes the IFST integrals into sums over 3-D grid points. The GIST method again uses MD simulations to generate appropriate sampling of water molecules across the protein surface. Following the MD simulations, the GIST methodology can be applied to yield a detailed view of the energy and entropy of areas with high water density. Of particular note is that the GIST methods have been implemented into the freely available AmberTools software, facilitating application of the methodology[97].

While these methods are able to give a detailed thermodynamic view of water molecules, and may therefore hint at the ability for a water molecule to be displaced, none of these methods are capable of predicting the displacing group without additional calculations. For example, Haider and Huggins used IFST in combination with multiple-copy simultaneous search to identify favorably displaced waters of HSP90 and to predict which functional groups were likely to displace them[98]. Cosolvent simulations are a promising means of predicting conserved or displaced waters, as they inherently account for the effects of solvation when

considering binding of probe molecules. Alvarez-Garcia and Barril utilized the MDmix method to predict water displacement of HSP90 and HIV protease[99]. Using a 20% concentration of either ethanol or acetamide in water, they performed 3 20ns MD simulations for each target. The predicted binding affinities of either probe or water molecules are then compared using the inverse of the Boltzmann equation, given in Equation 4. In order to assess the predictive power of the MDmix method, the resulting probe occupancies were compared to the experimentally known preferred interaction types, taken from structures of ligand-bound HSP90 and HIV protease. For the targets tested, Alvarez-Garcia and Barril found reasonable overlap between the MDmix probe occupancies and the functional groups of known ligands. For HSP90, 13 of 20 ligand interactions were replicated by MDmix, compared to 18 of 29 for HIV protease. This inability to reproduce all known interaction types found in HIV protease and HSP90 inhibitors may be partially due to the limited set of probes used. The authors then evaluated the ability of MDmix to predict water displacement by comparing the calculated binding affinities with the proportion of crystal structures containing a water molecule at a specific site. While the authors observed a fairly good correlation between their predicted displacement score and the experimentally observed displacement (R^2=0.72 for HSP90), this manner of comparison is fundamentally flawed. Using the frequency of observing a water molecule in a crystal structure as a substitute measure of its ease of displacement is incorrect. Crystal structures are not solved with the intention of systematically testing if every water molecule is potentially displaceable. Moreover, crystal structures are frequently solved containing a related series of ligands, which will bias the observed frequency of displacement for a given water molecule. For example, if only one ligand places an R group at a particular site and displaces a water molecule, while all other ligands are focused around a central core, the metric used by the MDmix group would indicate that the water molecule at the edge site is difficult to displace while those at the center of the group of ligands are easily displaced. It is possible though, that the binding affinity of the waters is exactly equal, or that those in the center of the ligands are actually more tightly bound and have a higher binding affinity. The observed linear correlation between experimental displacement and predicted displacement is therefore just coincidence, and is highly variable depending on the available crystal structures.

1.8 Overview and Aims of Thesis

While several groups have been instrumental in the development of cosolvent MD techniques, there are still a number of limitations preventing their widespread application. The primary aim of this thesis is to address these limitations, so that cosolvent MD simulations may be prospectively applied to identify binding sites and assist in structure-based drug design. For example, existing analysis methods require a great deal of manual inspection to interpret the results of the simulations. To address this, we have introduced MixMD Probeview, which automates the identification and ranking of potential binding sites from cosolvent simulations. Described in Chapter 6, MixMD Probeview is available as a plugin for the free, open-source version of PyMOL[100]. MixMD Probeview shifts the analysis of MixMD occupancy maps from primarily qualitative ranking to a more quantitative analysis of overall occupancy. This enables binding sites to be clearly distinguished from other easily desolvated sites.

Furthermore, there are no freely available protocols for converting cosolvent simulation results into pharmacophore models for prospective use in screening ligands. Using ABL kinase as a test system, we have developed a series of scripts, described in Chapter 5, which convert occupancy maps into pharmacophore features in the format required for virtual screening with the program MOE[30]. This enables cosolvent simulation results to be utilized in a prospective manner. Additionally, we have characterized occupancy levels during cosolvent simulations for non-displaceable water sites. The incorrect treatment of binding-site water molecules is a major source of error in predicting protein-ligand interactions. Using the analysis described in Chapter 4 allows for conserved water molecules to be identified and accounted for in subsequent structure-based drug design efforts.

Lastly, we have examined the potential of accelerated molecule dynamics to enhance sampling in combination with MixMD. Due to the time consuming nature of molecular dynamics simulations, the extent of conformational changes which may be studied are limited. As detailed in Appendix B, accelerated MixMD allows for faster convergence and promotes

greater conformational sampling, thereby extending the number of systems that MixMD can be applied to. Altogether, these developments to MixMD enhance its predictive ability and facilitate application to a variety of structure-based drug design endeavors.

As a secondary focus, we have carried out two additional studies. These are described in Chapters 2 and 3, while the remainder of this thesis focuses on the development of MixMD. Chapter 2 details an epidemiological study that I contributed to in fulfillment of the clinical research component of the Translational Research Education Certificate through the Michigan Institute for Clinical & Health Research. This study was aimed at understanding the transmission and sequence variation of CTX-M-type β-lactamases. Chapter 3 describes the use of traditional MD simulations to understand the dynamics of NSD1. Existing crystal structures of NSD1 have the important post-SET loop in an autoinhibitory position. MD simulations allow for the predicted motions of this loop to be analyzed, yielding insight into its conformational behavior in solution.

Chapter 2. Detection and Sequencing of CTX-M β-lactamases in Clinical *E. coli* Isolates

This chapter has been adapted from the following publication:

Graham, S.E., Zhang, L. Ali, I., Cho, Y.K., Ismail, M.D., Carlson, H.A., Foxman, B. Prevalence of CTX-M extended-spectrum beta-lactamases and sequence type 131 in Korean blood, urine, and rectal *Escherichia coli* isolates. Infection, Genetics and Evolution. 2016. 41:292-295

2.1 Abstract

A high proportion of extended-spectrum beta-lactamase (ESBL) producing Escherichia coli are of the ST131 lineage, but there are few estimates of ST131 prevalence among ESBL-negative E. coli. Without this information, it is difficult to evaluate the contribution of the ST131 lineage to the emergence and spread of ESBL E. coli. A total of 1,658 E. coli isolates were collected at Gachon University Gil Medical Center in Korea from 2006 to 2008. The antibiotic resistance profile was determined for all isolates, and ESBL-positive isolates were screened for the presence of CTX-M-type ESBLs. All ESBL-positive (n=84) and a representative sample of ESBL-negative (n=100) isolates were screened for O25b-ST131 using a PCR-based assay. The isolates were further classified on the basis of fumC and fimH types, which allowed for a comparison of the two typing methods. 5.7% of isolates were ESBL-positive, 87% of which contained CTX-M-type ESBLs. There was no significant difference in the prevalence of ST131 between ESBL-positive and –negative groups; 14% of ESBL-positive isolates and 9% of tested ESBL-negative isolates were ST131 by CH-typing. ST131-positive isolates harbored CTX-M-1-group ESBLs (including CTX-M-15) more frequently than other CTX-M types, and exhibited greater levels of antibiotic resistance than non-ST131 isolates. Furthermore, a number of

isolates identified as O25b-ST131 by PCR corresponded to non-ST131 sequence types by CH-typing, emphasizing the need to consider the testing method when comparing reported prevalences of ST131.

2.2 Introduction

Escherichia coli is the most common cause of urinary tract infections (UTI) and a frequent cause of bloodstream infections. UTI treatment is increasingly complicated due to the spread of antibiotic resistant organisms. Of particular concern are the extended-spectrum beta-lactamases (ESBL), which are resistant to penicillins and oxyimino-cephalosporins [101]. The CTX-M group of ESBLs is currently the dominant type of ESBL observed in *E. coli* [102]. CTX-M-containing isolates are often multidrug-resistant, especially to the UTI treatments of choice: fluoroquinolones and trimethoprim-sulfamethoxazole [103-106].

The increasing prevalence of CTX-M type ESBLs among *E. coli* isolates, specifically type CTX-M-15, is attributed to the spread of sequence type 131 (ST131) [107-109]. Due to the initial association with ESBLs, the majority of ST131 studies have described isolates that have ESBL resistance, or compare matched sets of resistant and susceptible isolates [110]. In order to understand the role that ST131 has played in the spread of ESBLs, it is necessary to better estimate the prevalence of ST131 among ESBL-negative isolates.

As sequence type assignment using multi-locus sequence typing (MLST) is time-consuming and expensive, several techniques have been developed which aim to identify ST131 isolates using PCR and/or sequencing of selected genes[110]. For example, Clermont et al. developed a PCR-based assay for an O25b-ST131-specific polymorphism in the *pabB* gene [111]. Weissman and coworkers proposed the use of CH-typing, determined by sequencing of internal fragments of *fimH* and *fumC,* to identify sequence types and partition them into subgroups [112]. While a comparison of the three MLST schemes and the corresponding CH-types is available, most studies make use of a single typing method for reasons of practicality [113]. In order to

understand the effect of different typing methods on the observed prevalence of ST131 among clinical isolates, we used both a PCR-based assay for O25b-ST131 and CH-typing.

Herein, we present the prevalence of ST131 among all ESBL-positive, and a random sample of ESBL-negative blood, urine, and rectal *E. coli* isolates obtained from the Gil Medical Center in Korea between 2006 and 2008. Further, we compare the antibiotic resistance profiles and presence of CTX-M among ST131 and non-ST131 isolates.

2.3 Methods

Bacterial Strains

The entire 2006-2008 collection from Gachon University Gil Medical Center in Korea consisted of 94 ESBL-positive isolates (76 urinary, 17 blood, and 1 rectal) and 1564 ESBL-negative isolates (707 urinary, 373 blood, and 484 rectal) as described previously [114]. For the current study, we included all viable ESBL-positive isolates (66 urinary, 17 blood, and 1 rectal) and a random sample of ESBL-negative isolates (using the RAND function in Excel) to represent the source distribution in the original collection; 24 blood isolates, 45 urinary isolates, and 31 rectal isolates.

The collection includes four categories of *E. coli* isolates: 1) All *E. coli* positive blood cultures from inpatients with bacteremia during January 2006 to December 2008. 2) All *E. coli* positive urinary cultures from patients with urinary tract infections (UTIs), defined as the presence of greater than 10^5 CFUs/mL bacterial growth collected from a midstream specimen between December 2006 and December 2008. 3) *E. coli* urinary cultures from asymptomatic UTI patients, using the same definition and dates as in (2). 4) *E. coli* rectal isolates from healthy individuals who attended the Health Promotion Center of Gil Medical Center between September and December 2007. Isolates were frozen at -80°C in Glycerol/Luria Broth (1:1) until further testing for this study.

Susceptibility Testing

Rectal isolates were initially screened for *E. coli* with UriSelect media (Bio-Rad). Species were identified using the VITEK system (bioMérieux), and all 1,658 isolates were screened for susceptibility to amikacin, ampicillin, cefotaxime, ciprofloxacin, gentamicin, imipenem, and trimethoprim-sulfamethoxazole using the disk diffusion method. ESBL producing isolates were identified using the microdilution method. Results were classified according to the CLSI criteria (2010).

PCR Detection of CTX-M ESBL

Primers designed to amplify all known CTX-M variants were used [115]. PCR was carried out in 25 µL volumes using 12.5 µL GoTaq DNA polymerase (Promega), 9.5 µL water, 2.5 µL template DNA (extracted by boiling lysis), and CTX-M forward and reverse primers to a final concentration of 200 nM using the published conditions[115]. The PCR products were run on agarose gel, the resulting bands were purified (QIAquick gel extraction kit, QIAGEN), and sequenced (University of Michigan DNA Sequencing Core) using forward primers to determine the CTX-M group present. CTX-M types were determined using NCBI BLAST to compare known types with the sequencing results[116]. Statistical analysis was done using SPSS (IBM, version 22) and OpenEpi. Significance was determined using the Chi-square test.

PCR Detection of O25b-ST131

Isolates were screened for the presence of O25b-ST131 *E. coli* using primers identified by Clermont et al.[111]. PCR was carried out in 25 µL volumes, with 12.5 µL GoTaq DNA polymerase (Promega), 5 µL water, 2.5 µL template DNA (extracted using QIAcube, QIAGEN), and 2.5 µL each of 10 µM *pab*B and *trp*A. The PCR reaction was performed under the following conditions: initial denaturation at 94°C for 4 minutes, 30 cycles of 5 seconds at 94°C, 10 seconds at 65°C, 1 minute at 72°C, followed by a final extension at 72°C for 5 minutes. Results were visualized on agarose gels. Known O25b-ST131 and K-12 *E. coli* were used as positive and negative controls, respectively. A subset of the samples were previously typed by MLST and served as additional controls [114].

fumC/fimH Typing

CH-typing was performed using the published conditions[112]. PCR reactions were carried out with 12.5 µL GoTaq DNA polymerase (Promega), 2.5 µL forward and reverse primers, 2.5 µL template DNA (extracted using QIAcube, QIAGEN), and 2.5 µL water. The PCR products were run on agarose gels to confirm a band of the expected size. Subsequently, the products were purified (QIAquick PCR purification kit, QIAGEN) and sequenced with forward and reverse primers (University of Michigan DNA Sequencing Core). The resulting sequences were trimmed and aligned using CodonCode. *fumC* and *fimH* types were assigned using the available web-services (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli and https://cge.cbs.dtu.dk/services/FimTyper-1.0/) and NCBI BLAST, and compared to published CH-types [112, 117]. Significance was determined using the Chi-square test or Fisher's exact test as appropriate.

2.4 Results and Discussion

Prevalence of ESBL and ST131 among Tested Isolates

Overall, 5.7 % of the collection's isolates had the ESBL phenotype. The ESBL phenotype was significantly more common among urinary isolates than blood isolates (9.7% versus 4.4%, p=0.0014) and least common among rectal isolates (0.2%). These results are consistent with a previous report from Korea [118]. Based on CH-types, there was no significant difference in the prevalence of ST131 by ESBL-phenotype or isolate source. The prevalence of ST131 was 14.3% (12/84) among ESBL-positive isolates, and 9% (9/100) among ESBL-negative isolates (p=0.26). These values are near the range reported in previous Korean studies of ESBL-positive *E. coli* (19.7% to 36.2%, by MLST) [108, 119-121].

Within the ESBL-positive isolates, 15.2% (10/66) of urine isolates and 11.8% (2/17) of blood isolates were ST131 by CH-typing (p=1.0). Within the ESBL-negative isolates, 15.6% (7/45) of urine isolates and 4.2% (1/24) blood isolates were ST131 (p=0.31). Only 1 rectal isolate in the tested subset was ST131 Figure 2.1. With regard to CH-types, 9/12 ESBL-positive isolates had CH-type 40-30, 1 was type 40-41, 1 was type 40-29, and 1 was *fumC* type 40 and *fimH* null. To the best of our knowledge, CH-type 40-29 has not been previously described as ST131, however *fimH* type 27 has been found in ST131 isolates [122]. *fimH* type 29 differs from *fimH* type 27 by only 1 base pair within the CH-typing region; therefore the isolate was assumed to be ST131. Within the ESBL-negative isolates, 5/9 were CH-type 40-30 and 4 were type 40-41.

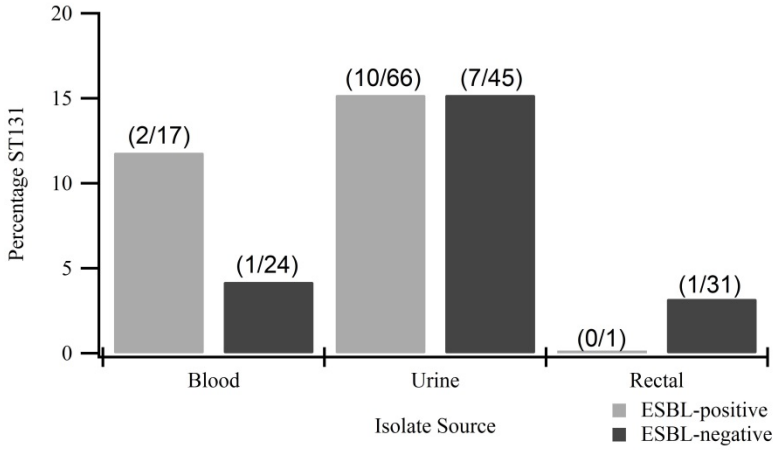


Figure 2.1: Prevalence of ST131 by source and ESBL phenotype within *Escherichia coli* isolates positive for ESBL (n=84) and representative sample of non-ESBL (n=100) from the 2006-2008 collection of Gachon University Gil Medical Center in Korea[120]

Association between ST131 and CTX-M

Almost all of the ESBL-positive isolates carried CTX-M (87%); there was no significant difference in prevalence by source – although the one ESBL positive rectal isolate did not carry CTX-M. ST131 positive isolates contained CTX-M-1 group enzymes (including CTX-M-15) more frequently than other ESBL isolates (6/9 or 67% versus 32/64 or 50%). This was not true for CTX-M-9 group enzymes which were less frequent among ST131 isolates (3/9 or 33% vs 33/64 or 52%; p=0.53). A summary of results is shown in Table 2.1.

	CTX-M Positive	
Source (n)	Percentage	n
ESBL positive (84)	86.9%	73
Blood (17)	94.1%	16
Rectal (1)	0.0%	0
Urinary (66)	86.4%	57

Source (n)	CTX-M-1 group	CTX-M-9 group
CTX-M positive	52.1%	49.3%
Blood[+] (16)	15	2
Urinary (57)	23	34
ST131		
Positive (9)	66.7%	33.3%
Negative[+] (64)	50.0%	51.6%

[+]1 isolate had both CTX-M-1 and CTX-M-9 group ESBLs

Table 2.1: Presence of CTX-M by source and sequence type among extended-spectrum beta-lactamase (ESBL) positive *Escherichia coli* isolates from the 2006-2008 collection of Gachon University Gil Medical Center in Korea[120].

Association between ST131 and Antibiotic Resistance

In general, ST131 blood and urine isolates had higher levels of resistance to most antibiotics than non-ST131 isolates, as shown in Figure 2.. ST131 isolates were less frequently resistant than non-ST131 isolates to trimethoprim-sulfamethoxazole (7/20 vs. 65/132, p=0.24). Previous studies have shown conflicting results in regards to ST131's resistance to trimethoprim-sulfamethoxazole[123, 124]. Similar to previous studies, ST131 isolates in the present study were more resistant to amikacin than non-ST131 isolates[123]. Whole genome sequencing of 104 *E. coli* ST131 identified from a collection of 1,908 consecutive single-patient *E. coli* from the United States and Germany suggests that fluoroquinolone resistance is primarily confined

to the H30-R subclone of ST131 [125]. This was also true in a 2012-2013 Korean study of 268 consecutive *E. coli* urinary and blood isolates from 21 Korean hospitals where 21% were ST131: all of the ST131 of the H30 subclone were resistant to ciprofloxacin (48/48) compared to 50% of the non-H30 (4/8) [126]. In the current study, 13/14 CH-type 40-30 isolates were resistant to ciprofloxacin compared to 1/5 CH-type 40-41.

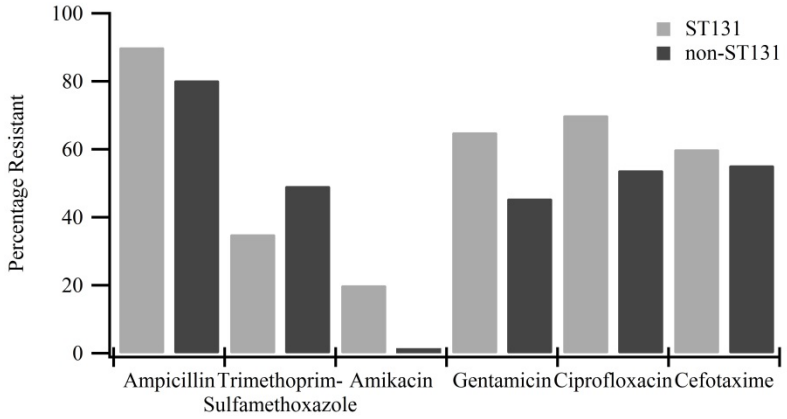


Figure 2.2: Antibiotic resistance by the ST131 phenotype among blood and urine *Escherichia coli* isolates positive for ESBL (n=83) and representative sample of non-ESBL (n=69) from the 2006-2008 collection of Gachon University Gil Medical Center in Korea[120].

Comparison between ST131 Assignment Methods

Isolates in this study were initially screened for the presence of O25b-ST131 using the method of Clermont et al. Although the prevalence of ST131 among ESBL-positive isolates (18/84, 21%) was seemingly in line with previous studies, the majority of ESBL-negative isolates (83/100, 83%) were classified as O25b-ST131. Testing was repeated multiple times by different individuals using fresh reagents and MLST-typed controls. Comparison with CH-types using the method of Weissman et al. revealed that only 9/100 of the ESBL-negative isolates were ST131. Within the ESBL-positive isolates, 38.9% (7/18) of isolates identified by PCR-based typing as O25b-ST131 were assigned CH-types corresponding to sequence types other than ST131. For example, 4 ESBL-positive isolates were positive by PCR for O25b-ST131 but corresponded to ST95 by CH-typing. This was also seen in a previous study using the PCR-based assay, in which

ST95 isolates were misclassified as ST131 [127]. A comparison of the results of the two-typing methods for all isolates combined is shown in **Table 2.2**.

	*fumC/fim*H Type						
pabB PCR Results	40-30 (ST131)	40-41 (ST131)	40-29 (ST131)	40-0 (ST131)	38-27 (ST 95)	Other ST131-Complex	Other Non-ST131
O25b-ST131	13	2	1	1	7	33	44
Non-O25b-ST131	1	3	0	0	0	45	34

*ST131 complex was defined as isolates with *fum*C types included in the ST131 complex by Achtman MLST, not including ST131 itself

Table 2.2: ST131 assignment using *pabB* compared with assignment using *fumC/fimH*. Extended-spectrum beta lactamase (ESBL) positive *Escherichia coli* (n=84) and a sample of 100 ESBL negative *E. coli* from the 2006-2008 collection of Gachon University Gil Medical Center in Korea[120].

The potential for false-positives using ST131 classification methods based on individual genes has been previously recognized[113]. In the present study, we found the results of CH-typing to align more closely with previously reported values and antibiotic resistance phenotypes of ST131 isolates than the PCR-based method of Clermont et al. CH-typing is also advantageous as it is able to identify alternative sequence types by linking CH-types with sequence types. For these reasons, we suggest the use of the CH-typing method rather than methods based solely on PCR when alternatives to full MLST are desired.

2.5 Conclusions

Previous studies of ST131 *E. coli* have primarily focused on resistant isolates, making it difficult to determine the role of the ST131 lineage itself in the dissemination of ESBLs. In the present study, we observed a similar prevalence of ST131 in both ESBL-positive and –negative isolates, particularly among urinary isolates. Further partitioning by CH-typing allowed for a comparison of resistance within the ST131 group, and found CH-type 40-30 isolates to be more frequently resistant to ciprofloxacin and slightly more frequent among ESBL-positive isolates than CH-type 40-41.

We characterized isolates collected during a period in which ST131 is thought to have been rapidly expanding. These data are especially pertinent, as they offer insight into the spread and evolution of ST131 over time; information essential for predicting future expansion of ST131 and understanding ST131's potential role in the spread of additional antibiotic resistance genes. However, there are a few limitations. Isolates were collected at only one study site, which limits the generalizability of our data. Data on antibiotic usage among participants is not available, and so differences in antibiotic exposure and selection for resistant *E. coli* cannot be accounted for. Nevertheless, our results indicate that the difference in prevalence of ST131 *E. coli* between ESBL-positive and ESBL-negative isolates during this time period was not significant, and that variation in the resistance phenotype within the ST131 group can be identified by CH-types.

Chapter 3. Dynamic Behavior of the Post-SET Loop Region of NSD1

This chapter has been adapted from the following publication:

Graham, S.E., Tweedy, S.E., Carlson, H.A. Dynamic behavior of the post-SET loop region of NSD1: Implications for histone binding and drug development. Protein Science. 2016, 25(5):1021-1029.

3.1 Abstract

NSD1 is a SET-domain histone methyltransferase that methylates lysine 36 of histone 3. In the crystal structure of NSD1, the post-SET loop is in an autoinhibitory position that blocks binding of the histone peptide as well as the entrance to the lysine-binding channel. The conformational dynamics preceding histone binding and the mechanism by which the post-SET loop moves to accommodate the target lysine is currently unknown, although potential models have been proposed. Using molecular dynamics simulations, we have identified potential conformations of the post-SET loop differing from those of previous studies, as well as proposed a model of peptide-bound NSD1. Our simulations illustrate the dynamic behavior of the post-SET loop and the presence of a few distinct conformations. In every case, the post-SET loop remains in an autoinhibitory position blocking the peptide-binding cleft, suggesting that another interaction is required to optimally position NSD1 in an active conformation. This finding provides initial evidence for a mechanism by which NSD1 preferentially binds nucleosomal substrates.

3.2 Introduction

Histone methylation is an important genetic regulatory element, among other post-translational modifications. Methylation occurs at a number of different sites, including arginine, lysine, and histidine residues, and is catalyzed by the methyltransferase group of enzymes[128]. The histone lysine methyltransferase enzymes are responsible for catalyzing the transfer of a methyl group from s-adenosylmethionine (AdoMet) to the target lysine. Depending on the target residue and specific enzyme, the lysine may be mono-, di-, or tri-methylated. These enzymes feature a conserved SET domain composed of approximately 130 residues, and they often have similar pre- and post-SET motifs[129].

The NSD family of histone methyltransferases are responsible for mono- and dimethylation of lysine 36 on histone 3 (H3K36), and other targets have also been reported[130]. The NSD family of proteins have been implicated in several types of cancer, including acute myeloid leukemia, breast cancer, glioma, neuroblastoma, and multiple myeloma[131-134]. In the case of acute myeloid leukemia, methylation of H3K36 by NSD1 was shown to be essential to gene activation and leukemogenesis through the presence of NUP98-NSD1 fusion proteinswhich alter transcriptional regulation[131]. Histone methyltransferases are therefore a promising drug target for the treatment of these diseases, although development efforts are hampered by the difficulty in achieving specificity among a protein family characterized by their conserved catalytic domain.

One way to achieve ligand selectivity is to capitalize on differences in flexibility between potential binding targets[135]. The post-SET loop of SET-domain histone methyltransferases exhibits great conformational diversity. For example, NSD1, ASH1L, and SETD2 have all been crystallized with the post-SET loop in an autoinhibitory position (the protein's conformation is such that the entrance to the active site is physically blocked), the structures of DIM-5, NSD3 (PDB: 4YZ8, unpublished data), and MLL show a disordered loop, while the structure of G9a has the loop in an alpha-helical conformation [136-141]. Recent structural studies of the catalytic

domain of NSD1 and the homologous protein ASH1L indicate two potential autoinhibitory mechanisms observed in the post-SET loop, as shown in **Figure 3.1**[137, 138]. A similar autoinhibitory loop position is also observed in the structure of SETD2[139]. In either case, a conformational change must occur in this loop region in order for the lysine to enter the substrate-binding channel. Previous studies have primarily focused on the structure of NSD1 after manual insertion of a peptide rather than free-dynamics of the loop. The dynamics of the post-SET loop are significant, as the loop must move prior to peptide binding or its movement must be induced upon interaction with the substrate. In addition, mutagenesis studies on the corresponding region of ASH1L have suggested that this loop region is not merely a blockade to peptide binding, but rather plays a more complex role in enzymatic activity[142]. As ligands targeting this region of NSD1 would have to bind prior to the peptide binding in order to be effective, it is necessary to understand the dynamics of this loop region.

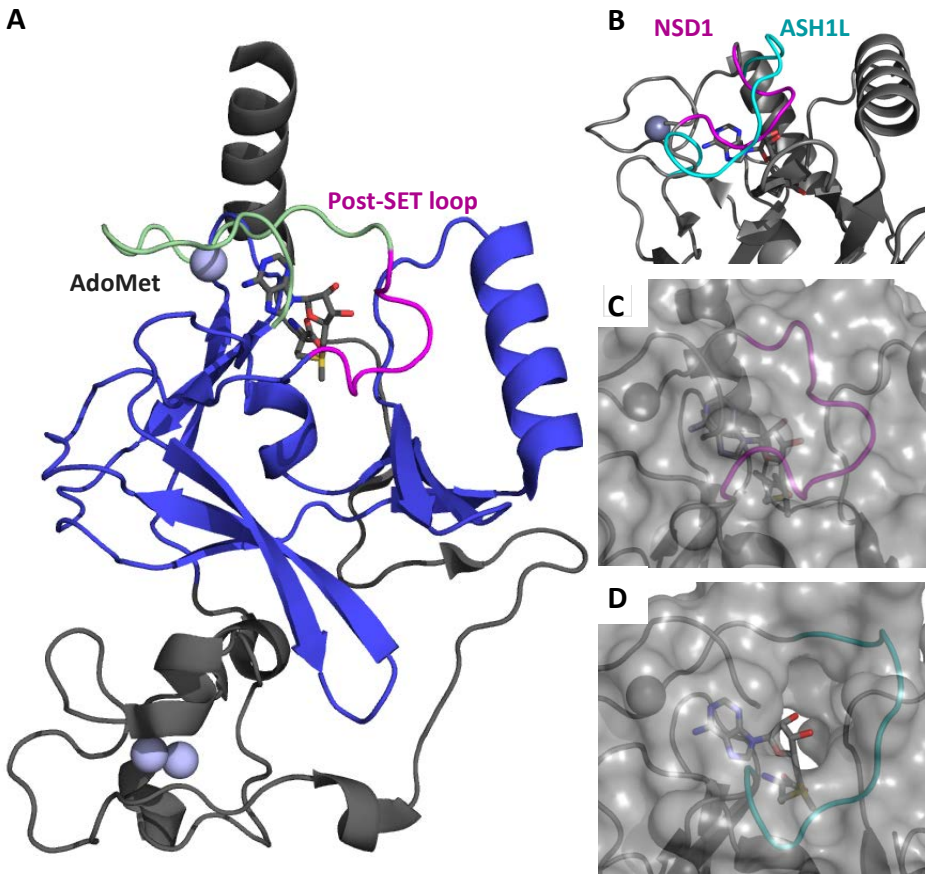


Figure 3.1: Crystal structures of NSD1 and ASH1L
A) The structure of the catalytic domain of NSD1 is shown (PDB: 3OOI). The SET domain is shown in blue, the post-SET loop is shown in magenta, and the post-SET domain is shown in green. The Zinc ions are shown as gray spheres. B) The post-SET loop region of ASH1L is shown in cyan in comparison with the post-SET loop region of NSD1 in magenta. For clarity, only the loop region of ASH1L is shown. C) Surface representation of the post-SET loop of NSD1 which shows the lysine-binding channel to AdoMet obstructed by the post-SET loop. D) Surface representation of the post-SET loop of ASH1L, showing a cavity not found in the crystal structure of NSD1

Few studies thus far have analyzed the dynamic behavior of NSD1, so it is unclear what conformations exist prior to histone binding. Qiao and coworkers performed a 2 ns MD (Molecular Dynamics) simulation in order to examine the conformational variability of the post-SET loop region[137]. Although they were able to see "modest conformational changes", the length of their simulation was relatively short and so they may not have captured other potential conformations. In addition, they used docking and energy minimization of the

nucleosome against NSD1 to position H3K36 into the substrate binding channel. In their procedure, the majority of the nucleosome was held rigid while the histone tail was allowed to move. In the resulting model, the nucleosomal DNA contacts the post-SET loop, which may play a role in stabilizing the active conformation[137]. Work by the di Luccio group also utilized MD simulations to analyze a peptide-bound model of the post-SET loop[143, 144]. However, their computational work has serious technical flaws; namely the use of a 1-ps time step. Typical time steps for molecular dynamics simulations are on the order of 1-2 fs. The 1-ps time step is orders of magnitude too large, which leads to instabilities in the forces and poor sampling in the calculated motion of the protein[145]. Additionally, they minimized their structures after simulation, so it is unclear what conformations were observed during the simulations themselves.

While the previous studies give some insight into the conformational diversity observed in the post-SET loop, they have not sufficiently examined the potential conformations of NSD1 without the peptide bound which may be an important state for developing inhibitors of methyl transfer nor in the presence of the H3 peptide to better understand the substrate-bound state of NSD1. In order to determine which conformations this loop may adopt, we have performed MD simulations of NSD1 with the post-SET loop as observed in the crystal structure as well as simulations with the loop modeled to match that seen in homologous structures.

3.3 Methods

In the present study, we utilized Molecular Dynamics simulations to examine three possible starting conformations of the NSD1 post-SET loop: 1) a peptide-bound model, 2) the position observed in the crystal structure ("closed-inactive"), and 3) the position observed in the homologous protein ASH1L ("open-inactive"). In the closed-inactive structure, the post-SET loop folds more compactly towards the AdoMet cofactor in contrast to the more-relaxed position observed in the open-inactive simulations.

Peptide-Bound Conformation

No peptide-bound crystal structures of NSD1 were available, so we first created the model of the histone H3 peptide (residues 33-37) bound to NSD1-AdoMet based on the structurally homologous SET domain of the H3K9 methyltransferase GLP (PDB: 3SW9)[146]. This was done using the program MOE to manually adjust the post-SET loop of NSD1 to mirror the positioning observed in the structure of GLP and to insert the corresponding peptide[30]. Due to the large number of rotatable bonds in a peptide, modelling the peptide-bound conformation using a homologous structure produces a more reliable model than docking the peptide into the structure. While previous studies have shown that a post-SET extension of NSD1 is required for nucleosome binding, no crystal structures containing the both SET domain and this region exist. Although hypothetical models have been proposed, modeling this region in our simulations would introduce another level of uncertainty; therefore we have chosen to use the construct of NSD1 found in the crystal structure (PDB: 3OOI)[137, 147].

Inactive Conformations

Second, we sought to enumerate potential conformations of the post-SET loop that may exist prior to binding of the nucleosome. To do so, we carried out two sets of simulations based on binary NSD1-AdoMet. The first, which we have termed the closed-inactive position, started from the crystal structure of NSD1 (PDB: 3OOI)[137]. During our simulations, we observed a subset of the closed-inactive models transitioning to a more relaxed conformation resembling the autoinhibitory loop position seen in ASH1L. In order to determine if this is a stable conformation, we also modeled the post-SET loop to match the position seen in the crystal structure of ASH1L, termed the open-inactive position(PDB: 3OPE)[138]. This was done using the secondary structure matching utility within the program *Coot* to align the corresponding post-SET loop residues of NSD1 to those of ASH1L[148].

Computational Setup

Parameter files for the simulations were prepared using the tleap utility in AmberTools with the AMBER FF99SB force field[149]. The parameters for the AdoMet cofactor were created

using the antechamber utility in AmberTools with GAFF atom types and the AM1-BCC charge model[150, 151]. Protonation states were assigned using PROPKA[152, 153]. In the case of the peptide-bound model, the target lysine was modeled as neutral in order to more closely mimic the transition state preceding methylation. This choice was based on the computational study by Zhang and Bruice which indicated the presence of a neutral lysine that is deprotonated through a water channel prior to methylation in related SET-domain lysine methyltransferases[154]. The systems were solvated with TIP3P water and sodium or chloride ions were added as necessary to achieve a net-neutral charge[155]. The systems were initially minimized for 7500 steps, followed by heating to 300 K over 80 ps with a 2-fs time step and restraints on the protein. The systems were then equilibrated at 300 K for 1.75 ns while the restraints were gradually released. The SHAKE algorithm was used to constrain bonds with hydrogens and the temperature was controlled with the Berendsen thermostat[156]. Ten separate production runs of 50 ns were completed for each of the three setups. All simulations were carried out using the sander and pmemd utilities in AMBER11[149]. The trajectories were analyzed using the ptraj and cpptraj utilities[157]. Visualization of trajectories was done using VMD and PyMOL[100, 158].

Movement Metric

In order to quantify the movements of the post-SET loop during the simulations of the inactive conformations, we selected two metrics that captured the observed rotation and translation of the loop. This allowed us to assess the extent of conformational variation in each trajectory. The first metric calculated was the average of two dihedrals, Gly213N-Gly213Cα-Gly213C-Asn214N and Gly104C-Asn214C-Asn214Cα-Asn214Cγ, which measured the rotation of the post-SET loop away from AdoMet. The second metric measured the distance between Cγ of Asn214 and the O3' atom of AdoMet. For ease of comparison, both metrics were normalized from 0 to 1 before plotting.

3.4 Results

Peptide-Bound Simulations

Over the course of our simulations, the post-SET loop relaxed into a stable conformation with minimal changes to the rest of the protein. RMSD measurements confirm this observation; the initial homology model deviated from the sampled trajectories with a minimal RMSD for the post-set loop region of approximately 3 Å. In comparison to the minimized and equilibrated homology model, RMSD values over the course of the simulation oscillated around 2 Å for both the post-SET loop and the entire protein. Each of the simulations of peptide-bound NSD1 exhibited similar behavior. In every case, the post-SET loop maintained a helical conformation for residues 208-215 (PDB: 3OOI numbering), as shown in **Figure 3.2**. This conformation allows the lysine of the peptide to remain adjacent to AdoMet in preparation for methyl transfer. Based on our simulations, this structure represents a potential model of NSD1 bound to the H3 peptide.

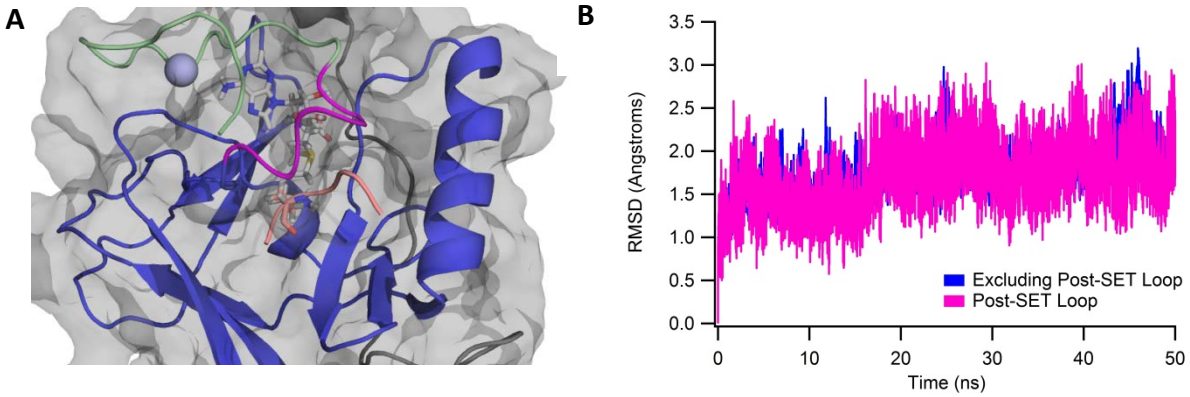


Figure 3.2: Model of Peptide-bound NSD1
A) Representative structure of peptide-bound NSD1. The H3 peptide is shown in salmon. This structure was chosen from a clustering of the final 10 ns of all peptide-bound runs, and is the representative structure from the highest occupancy cluster. B) Representative backbone RMSD plot calculated relative to the equilibrated peptide-bound model. As shown in magenta, the loop relaxed into a stable conformation with a relatively small RMSD to the starting structure, and remained stable, with minor oscillations of ±0.5Å around the average position.

Closed-Inactive Simulations

In contrast with the peptide-bound simulations, the post-SET loop region exhibited greater conformational variability during our simulations of NSD1-AdoMet. In the simulations starting from the closed-inactive structure, two of the runs adopted a conformation resembling the open-inactive structure, five of the runs sampled about the closed-inactive conformation, and three runs transitioned into a third autoinhibitory conformation.

The transition from the closed-inactive to the open-inactive structure was characterized by an outward movement of Asn214 (PDB: 3OOI numbering) and the simultaneous rotation of the adjacent post-SET loop region. The degree of conformational variation was assessed using the dihedral and distance metrics described in the methods section. The transition from closed-inactive to open-inactive can be seen in the first and second closed simulations, as shown in **Figure 3.3**A and B respectively, in which both the dihedral metric and distance metric increase. This corresponds to a movement of the post-SET loop away from the AdoMet cofactor and towards the nearby α-helix. This conformational change is confined to the region directly surrounding Asn214, with very little conformational change in the rest of the post-SET loop region. This outward movement of the Asn214 residue is not observed in the remainder of the closed-inactive simulations.

In three of the simulations (**Figure 3.3**F, I, J), we see a third conformational state, characterized by an increase in the distance metric without a corresponding increase in the dihedral metric. In these cases, Asn214 rotates in the opposite direction, away from both AdoMet and the alpha helix. In one case, this results in an enlargement of the upper portion of the peptide-binding cleft, although the lysine binding channel is still blocked. Interestingly, the cleft makes a flexing motion over time, initially forming a wider intermediate state, which then narrows somewhat at the end of the simulation. This is depicted in **Figure 3.3**K. This simulation was run for an additional 80 ns (for a total of 130ns) to determine if the loop would further transition into a conformation with the lysine binding channel exposed, however, the loop remained in the observed autoinhibitory position.

In the remainder of the simulations, shown in **Figure 3.3**C-E, G, and H, the loop remains in the closed-inactive position. In these cases, the post-SET loop primarily samples around the starting position and does not rotate outward.

Open-Inactive Simulations

There is much less conformational variation observed in the simulations that were initialized with the post-SET loop positioned like that of ASH1L. Using the same metric as the closed-inactive simulations, as shown in **Figure 3.4**, the loop maintains the initial structure while sampling around this position. The backbone RMSD fluctuates around 2 Å, while the backbone RMSD of the post-SET loop region relative to the homology model is approximately 2.5-3 Å over the course of the simulations. The open-inactive simulations were initiated with Asn214 pointing outwards towards the α-helix and away from AdoMet. In contrast to the closed simulations, Asn214 never rotates back towards AdoMet and always maintains this outward-facing position.

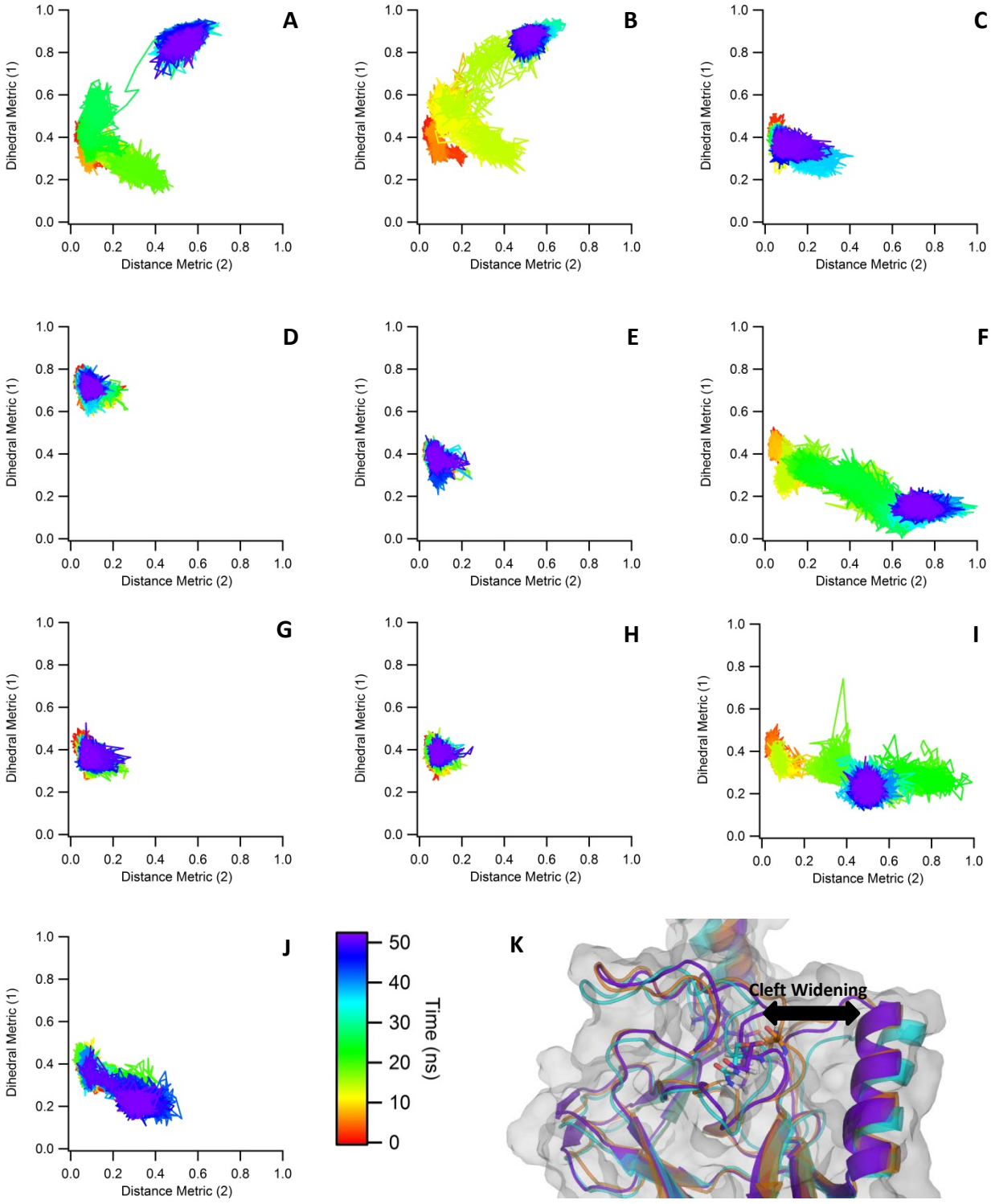

A
B
C
D
E
F
G
H
I
J
K
Dihedral Metric (1)
Distance Metric (2)
0.0
0.2
0.4
0.6
0.8
1.0
Time (ns)
0
10
20
30
40
50
Cleft Widening

Figure 3.3: Closed-Inactive Metrics
A-J) The metrics describing the movement of the post-SET loop in the closed-inactive simulations are shown. The average dihedrals are shown on the Y axis while the distance is shown in the X axis. All values shown are normalized for easier comparison. In five of the simulations, the post-SET loop samples around the starting position. In the remaining simulations, we observed a transition to two other distinct conformations. K) Initial, intermediate, and final conformations from trajectory "F" are shown, colored orange, cyan, and purple, respectively.

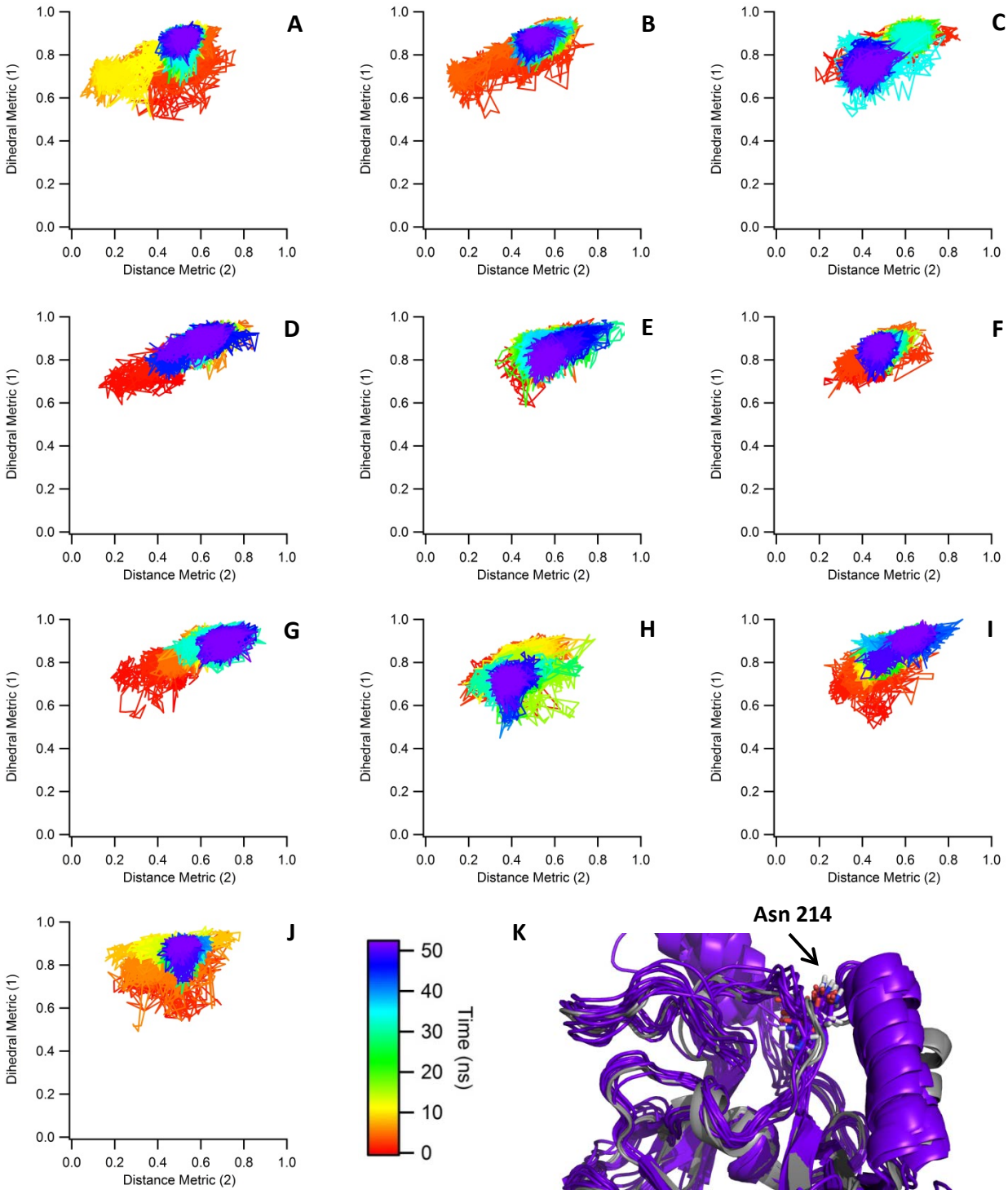

A
B
C
D
E
F
G
H
I
J
K
Dihedral Metric (1)
Distance Metric (2)
0.0
0.2
0.4
0.6
0.8
1.0
Time (ns)
0
10
20
30
40
50
Asn 214

Figure 3.4: Open-Inactive Metrics
A-J) The metrics describing the movement of the post-SET loop in the open-inactive simulations are shown. The average dihedrals are shown on the Y axis while the distance is shown in the X axis. All values shown are normalized for easier comparison. In all cases, the post-SET loop primarily samples about the starting conformation. K) Comparison between final structures from the "open-inactive" simulations (purple) and crystal structures of the homologous protein ASH1L (gray, PDB: 3OPE, 4YNM).

3.5 Discussion

In our simulations of NSD1, we witnessed several conformational changes. In the case of the NSD1-AdoMet simulations, we observe three conformations in total. The first is that found in the NSD1 crystal structure, PDB: 3OOI, which has the post-SET loop in an autoinhibitory position ("closed-inactive"). Second, we observed another autoinhibitory position resembling that of ASH1L, PDB: 3OPE ("open-inactive"). This conformation was observed in both the simulations beginning from the closed-inactive structure and in those starting from the open-inactive homology model. While we do not know how frequently this open conformation occurs in solution relative to the closed position, it was observed to be stable over the entire length of our simulations. Recent structural studies of the corresponding autoinhibitory loop of ASH1L showed similar conformational sampling within the loop region[142]. Interestingly, we observed a third conformation in which the post-SET loop moves upward, widening the peptide binding cleft. In previous studies by Qiao and coworkers, their docking simulations positioned the nucleosomal DNA against the portion of the loop blocking the lysine binding channel[137]. It is possible that the motion observed in our simulations is a precursor to histone binding, with the remainder of the post-SET loop moving upon contact with the nucleosomal DNA. While a shortened construct of NSD1 has been shown to methylate the H3K36 peptide *in vitro,* NSD1's preferred substrate is nucleosomal H3K36[130, 159]. We speculate that the necessary conformational change in the post-SET loop that opens the lysine binding channel and peptide binding cleft occurs spontaneously on occasion, thereby allowing for the methylation of peptide substrates, but is typically induced by the presence of the nucleosome via an induced-fit mechanism, thus explaining the preference for nucleosomal H3K36. Our study is limited by the fact that full-length crystal structures of NSD1 are unavailable, and so it is possible that there is another determinant of NSD1's specificity *in vivo*. Nevertheless, our simulations yield insight on the dynamics of the post-SET loop region.

The observed conformations may also be useful in the design of NSD1-specific small molecules. Currently known inhibitors of SET-domain histone methyltransferases function as

either competitive inhibitors of the substrate peptides or inhibitors of the cofactor AdoMet. For example, compounds such as BIX01294, E72, and UNC0321 have been found to block binding of the target peptide of H3 to the histone methyltransferase G9a, and Sinefungin derivatives have been developed that bind in place of AdoMet[139, 160-163]. Targeting histone methyltransferases via these mechanisms may be potentially difficult because of the need to develop specific inhibitors. Indeed, Nguyen et al. have studied the binding site similarity of SET-domain methyltransferases and found scaffolds that may bind to subsets of these[164]. They suggest that the substituents on these scaffolds may be changed to garner specificity. As an alternative to this, it may be possible to capitalize on the differing conformational changes between SET-domain methyltransferases, including those seen in our simulations of NSD1. As mentioned previously, the structure of the post-SET loop region varies within the histone methyltransferase family. By designing small molecules that stabilize the autoinhibitory loop rather than the cofactor or lysine binding site, researchers may be able to develop ligands with greater specificity for NSD1 over other SET-domain methyltransferases. The results presented here provide conformations that can be used for structure-based drug-design efforts to inhibit methylation by NSD1.

Chapter 4. Predicting Displaceable Water Sites Using Mixed-Solvent Molecular Dynamics

4.1 Abstract

Water molecules are an important factor in protein-ligand binding. Upon binding of a ligand with a protein's surface, waters can either be displaced by the ligand or may be conserved and possibly bridge interactions between the protein and ligand. Depending on the specific interactions made by the ligand, displacing waters can yield a gain in binding affinity. The extent to which binding affinity may increase is difficult to predict, as the favorable displacement of a water molecule is dependent on the site-specific interactions made by the water and the potential ligand. Several methods have been developed to predict the location of water sites on a protein's surface, but the majority of methods are not able to take into account both protein dynamics and the interactions made by specific functional groups. Mixed-solvent molecular dynamics (MixMD) is a cosolvent simulation technique that explicitly accounts for the interaction of both water and small molecule probes with a protein's surface, allowing for their direct competition. This method has previously been shown to identify both active and allosteric sites on a protein's surface. Using a test set of ten systems, we have developed a method using MixMD to identify conserved and displaceable water sites. Conserved sites can be determined by an occupancy-based metric to identify sites which are consistently occupied by water even in the presence of probe molecules. Conversely, displaceable water sites can be found by considering the sites which preferentially bind probe molecules. Furthermore, the inclusion of six probe types allows the MixMD method to predict which functional groups are capable of displacing which water sites. The MixMD method consistently identifies sites which are known to be conserved and predicts the favorable displacement of water sites that are known to be displaced upon ligand binding.

4.2 Introduction

Water molecules play an important role in protein-ligand interactions. The specific conservation or displacement of water molecules is a significant factor in molecular recognition[165], drug selectivity[166], and a ligand's binding affinity[88]. Upon ligand binding, waters at the binding interface must be displaced or participate in interactions between the protein and ligand. Waters at the binding interface fall into one of three categories: 1) waters which are always conserved, 2) waters which may be displaced by some ligands but not others, and 3) waters which are always displaced.

In the strategic design of ligands, scientists frequently try to increase a ligand's affinity by selectively displacing waters. It would be advantageous for researchers attempting this to have a means to predict whether a water site could be displaced, and whether this would lead to an increase in a ligand's affinity. To this end, a number of computational methods have been developed which attempt to predict the relative ease of displacement of a water site. For example, statistical methods such as AcquaAlta[86], Consolv[82], HINT/RANK[85], and Waterscore[83] utilize varying molecular descriptors such as crystallographic B-factors, number of hydrogen bonds, and descriptors of surrounding residues to analyze a hydration site. While these methods are relatively fast, they give predictive rates in the range of 50-70% depending on the method and test set used. Water prediction methods have also been incorporated into docking software. For example, the WaterDock methodology used with AutoDock Vina reported successful prediction of a water molecule's displacement in 75% of cases[167]. Alternatively, Monte Carlo simulations of water molecules may be performed to predict their locations and binding affinity, such as in the Just Add Water Molecules (JAWS) method[87, 88].

Since the specific interactions that determine whether a water molecule can be displaced or not are inherently site dependent, methods based on static structures may not accurately capture the variability among binding sites of different systems. Molecular dynamics-based methods are a promising alternative, as they are able to account for dynamics and interactions specific to each protein. For example, inhomogeneous fluid solvation theory

(IFST) provides a means of calculating binding energies, including enthalpic and entropic components, from molecular dynamics (MD) simulations[91, 92]. This method has been implemented into the WaterMap tool and successfully applied to a number of targets[93, 94, 168-171]. The SPAM method also utilizes molecular dynamics simulations to calculate the affinity of a water site by considering the probability distribution of the interaction energies of each water site with its surroundings[90].

While these methods are useful to analyze the energetics of individual water sites and predict their potential for displacement, they do not test the ability of specific functional groups to displace each site. In recent years, several cosolvent simulation techniques have been developed to map favorable interactions within a protein's binding site, including the MixMD, SILCS, and MDmix methods[52, 53, 63, 64, 67, 72, 73, 78, 79, 99]. In cosolvent MD simulations, a protein is initially immersed in a solution of small molecule probes and water. Following MD simulations during which the probes and water compete for binding to the protein's surface, the solvent occupancy can be calculated to identify locations on the protein's surface which preferentially interact with either the solvent probes or water. Furthermore, post-processing of the trajectories allows the binding affinity of each water site to be calculated[99]. While these methods are similar in their use of mixed solvents, each has methodological differences, such as the use of different probe molecules, whether individual probe molecules are run alone or in combination, and the use of restraints on protein and solvent atoms. The Mixed-Solvent Molecular Dynamics (MixMD) method has been developed by our group and previously shown to identify both active and allosteric sites[51]. In the present manuscript, we validate and extend the use of MixMD to map water sites and gauge their potential for displacement.

4.3 Methods

MD simulations

Ten systems were selected for the present test: Aldose Reductase (PDB:1ADS)[172], TEM-1 β-Lactamase (PDB:1ZG4)[173], β-Secretase (BACE, PDB:1W50)[174], Bromodomain Containing

Protein 4 (BRD4, PDB:2OSS)[175], Dihydrofolate Reductase (DHFR, PDB:1DG8)[176], Heat Shock Protein 90 (HSP90, PDB:1AH6)[177], Neuraminidase (PDB:4HZV)[178], Penicillin Binding Protein 4 (PBP-4, PDB:2EX2)[179], Penicillopepsin (PDB:3APP)[180], and Thrombin (PDB:3U69)[181]. These proteins were selected based on the criteria that they had apo crystal structures with better than 2 Å resolution and that each had multiple comparable ligand-bound structures in which water molecules were conserved, displaced, or selectively displaced relative to the apo structure. All crystallographic waters were removed prior to system setup. Hydrogens were added and side-chain positions were optimized using MolProbity[182]. Using a layered cosolvent approach, each protein was surrounded with a layer of probes (acetonitrile, isopropyl alcohol, *N*-methylacetamide, pyrimidine, or a methylammonium/acetate mix) followed by a layer of TIP3P water in a 5%/95% v/v ratio[73]. Simulations of the proteins in pure water were also done for comparison. Input files were prepared with tleap using the AMBER FF99SB force field and parameters developed by Ryde for NADP and NADPH[183-185]. Sodium or chloride ions were used to neutralize the systems, and ACE/NME caps were added to protein chains when appropriate. The systems were initially minimized with restraints on the protein for 5000 steps, followed by 2500 steps of minimization on the entire system. The systems were gradually heated at constant volume over 40,000 steps with a 2 fs timestep and restraints of 10 kcal/mol-Å^2 on the protein. After the systems had reached 300K, they were equilibrated at constant pressure for 1.75 ns as the restraints were gradually removed. Production runs were carried out for each system for 20 ns with the Andersen thermostat[186]. In total, 50 simulations were performed in AMBER12 for each protein; ten runs of 20 ns each per probe type[185]. Five simulations of 20 ns were completed for each of the systems in pure water. This provided a total of over 1 μs of total MD production for each protein.

Probe and Water Occupancy Calculation

The resulting trajectories were aligned using the AmberTools ptraj utility, and the occupancy of the probes and water during the last 10 ns of each simulation were calculated using a 0.5 Å grid over the entire solvent box[185]. To simplify further analysis, the resulting occupancies were normalized into σ units, using the equation:

$$\frac{x_i - \mu}{\sigma} \quad (1)$$

where x_i is the raw count at grid point i, μ is the mean occupancy over all grid points, and σ is the standard deviation across all grid occupancies. The occupancies can then be visualized in σ units, corresponding to the number of standard deviations above the mean occupancy (much like viewing electron density from crystal structures). Water and probe occupancy was visualized in PyMOL[100].

SPAM Binding Affinity

In order to compare the relative affinity of water for a specific site, we calculated the theoretical binding affinity for all water molecules using SPAM as implemented in AMBER14[90, 187]. The trajectories from the last 10 ns of each simulation (100 ns total per probe/system) were aligned and an in-house script was used to identify all water sites occupied at a level greater than the expected occupancy of bulk water. The interaction energies over every frame for each water site were then calculated with a cutoff of 10 Å using the SPAM routine in cpptraj[157]. An in-house script was used to convert the interaction energies into ΔG_{SPAM} following the given procedure[90]. ΔG_{SPAM} values given in the text are taken from the water-only simulations.

Water-site Conservation

To assess the ability of the MixMD method to find conserved water sites on a system-wide scale, we compared the water sites identified in the simulation with those found in comparable crystal structures. Comparable structures were identified using the Sequence Clusters from the Protein Data Bank at 95% sequence identity. This returns a list of crystal structures in the PDB at the specified similarity ranked by quality factor (based on resolution and R-value). The entries from this list (up to the top 99) with resolution better than 2.5 Å were selected for comparison. To identify hydration sites in each crystal structure, the structures were aligned using the wRMSD tool[188] and clusters of waters were identified using WatCH[189]. WatCH clusters water molecules using a 2.4 Å threshold to identify water molecules occupying the same region. Each cluster was considered to be a water site. The experimental

conservation of each water site was then calculated as the percentage of structures which had a water molecule within the cluster relative to the total number of structures. All reported percent conservation values in this manuscript refer to the percent conservation calculated from this analysis. It is important to note that observed experimental conservation will not necessarily be correlated with the displaceability of a water site. For example, multiple structures of a protein are frequently solved containing a series of related ligands, which may displace the same water molecule in every case. In addition, water molecules may be capable of being displaced, but ligands targeting that site may not yet have been developed (eg. waters on the edge of a binding site may be displaceable but current ligands do not extend that far). Nevertheless, this analysis gives some insight into the relative "ease of displacement" for each water site.

4.4 Results and Discussion

Predicting Water Displaceability

A method to predict water displacement would be most useful if it could definitively predict displacement using a clear-cut procedure. Therefore, we sought to create a set of guidelines with which to classify waters on the basis of MixMD results. The majority of current methods use predicted binding affinities to classify a water site as displaceable or conserved. Using the SPAM methodology, the theoretical affinity (ΔG_{SPAM}) for each water site was determined in the presence of water alone or with probes and water, which were then plotted as a histogram as shown in **Figure 4.1**. The simulations of water alone serve to identify the maximum number of water sites, as both conserved and displaceable sites will be identified. In simulations of water and probe, a decrease in the number of identified sites is expected, as water sites will be displaced by the probe molecules. As predicted, there is a large decrease in the number of waters with positive ΔG_{SPAM}, indicating their favorable displacement. However, a small decrease in the number of sites having weakly favorable (negative) ΔG_{SPAM} values is also seen, indicating that these sites were displaced by the probe molecules or their free energies became less favorable in the presence of the probes. While the SPAM-generated ΔG values

give insight into the energetics of each individual site, they do not provide an unambiguous way to classify sites as displaceable or not.

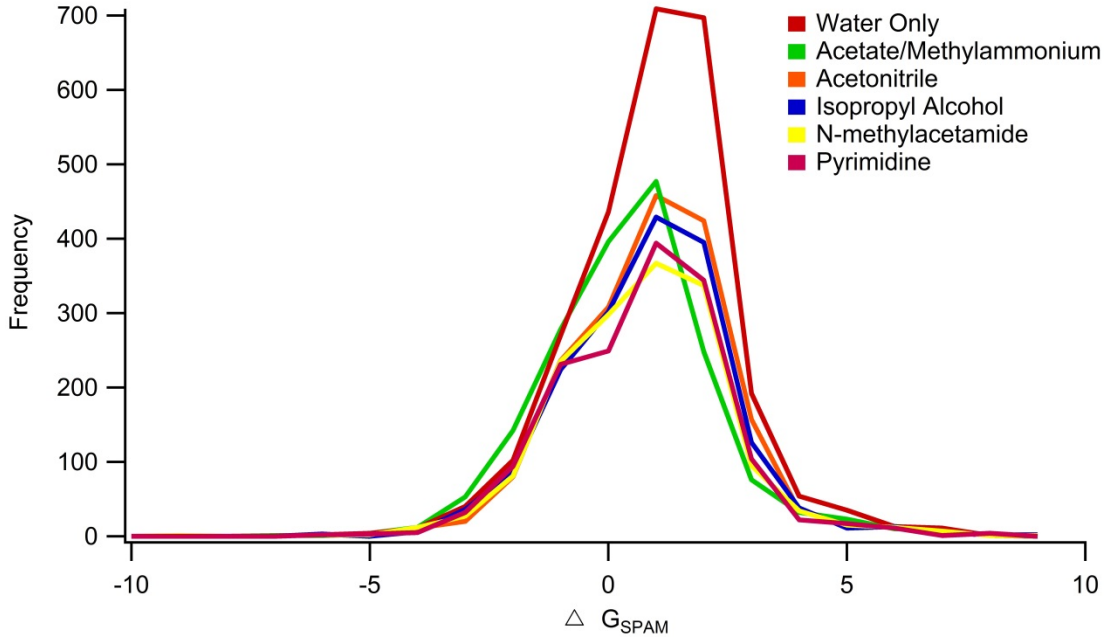
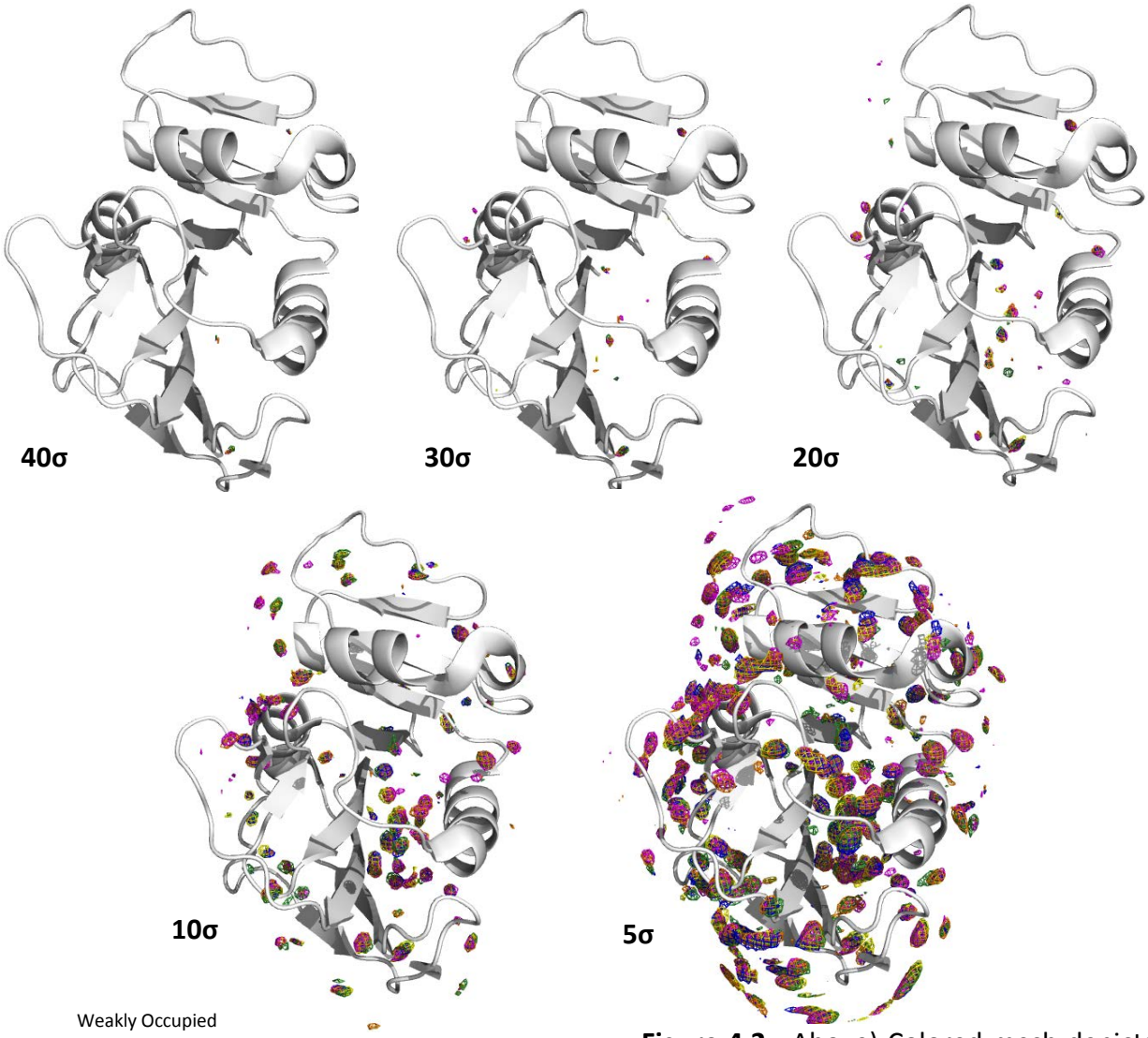


Figure 4.1: Histogram of SPAM-calculated binding affinities for water sites in each solvent type. ΔG_{SPAM} is binned in 1 kcal/mol increments. A decrease in the number of water-occupied sites is observed between the water-only (red) and water with probe simulations (colored lines), indicating the displacement of these sites by the probe molecules. Notably, there is a sharp decrease for water with positive ΔG_{SPAM}, but some waters with weakly favorable ΔG_{SPAM} are also displaced.

As an alternative to using a specific energy-based cutoff, the MixMD water occupancy can be visualized directly to identify sites which may be displaced. Since the MixMD simulations are performed with both small molecule probes and water, sites that would more favorably bind water over probe molecules (conserved sites) would have greater levels of water occupancy than sites which more favorably bind probes (displaceable sites). In order to determine what level of water occupancy throughout the simulations is characteristic of non-displaceable sites, each system was visualized in PyMol to identify water density within 10 Å of the MixMD identified binding hotspot of known ligands, as previously defined[51]. Local maxima were identified and their σ values (occupancies) binned to generate a histogram, as shown in **Figure 4.2**. When water occupancy is visualized at high σ values, only a few sites are observed. These sites are locations that are very frequently occupied by water despite the presence of

probe molecules, and are therefore considered to be non-displaceable. As σ values are decreased, sites that are less frequently occupied by water molecules are identified. Since these sites are not as frequently occupied by water when probe molecules are present, they are considered to be potentially displaceable. A cutoff of 20 σ was chosen to classify water sites as conserved or displaceable. In raw occupancy, this equates to a water molecule being at the same grid point in at least ~3% of all snapshots (note that a water site contains several grid points and the total occupancy of a water site is the sum over all of the associated grid points). The 20 σ value was found to be an ideal balance between identifying water sites which are conserved, while also not misclassifying sites which may be displaced. Contouring the occupancy at lower σ levels identifies sites which are known to be displaced; higher σ values miss sites that are known to be conserved. As distinct water sites will inherently have higher levels of expected occupancy than bulk water, we sought to describe the distribution of occupancies for only the local maxima within the active-site region. We have also included the cumulative probability distribution in **Figure 4.2**. The chosen 20 σ cutoff corresponds to a cumulative probability of ~0.84, and is therefore approximately 1 standard deviation above the mean occupancy for waters within the active site. In the sections that follow, results of this method for 10 systems are shown in order to demonstrate its ability to correctly predict water displacement, with examples of the method's ability to predict conserved and displaced waters shown for each system.



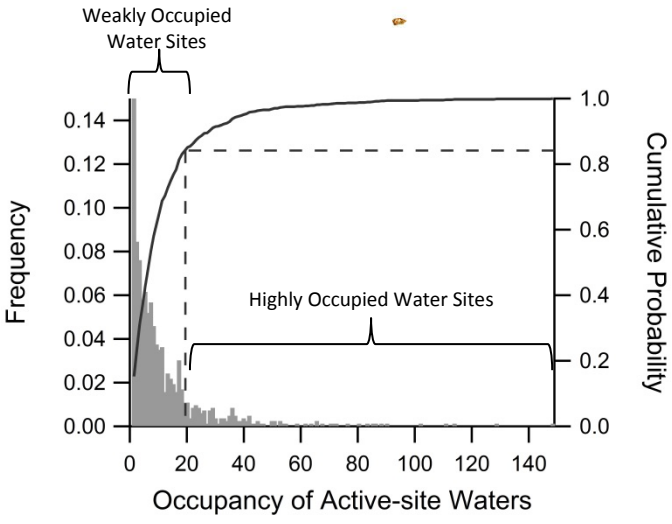


Figure 4.2: Above) Colored mesh depicts water occupancy from simulations of each probe and water mixture. At high occupancy levels, few water sites are identified. These are sites which are repeatedly occupied by water molecules even in the presence of probe molecules. Water sites that first appear at lower sigma values are less frequently occupied by water when probe molecules are present. Left) The distribution of normalized occupancies for water sites (local maxima) within the active-site region. Data is taken from water occupancy in the presence of all probe types and across all systems.

Comparing Site-Specific Binding Preferences

Aldose Reductase

Aldose reductase contains a cluster of buried water molecules which are conserved in 99% of all crystal structures, located at site A in **Figure 4.3**. Other water molecules in the binding pocket are displaced by ligands, including sites B and E, and are only conserved in 1-20% of crystal structures. Using the 20 σ cutoff, we correctly predict the cluster of conserved water sites and predict the displacement of water sites which are experimentally known to be displaceable. In **Figure 4.3**, the crystallographic water locations in the apo structure are shown as red spheres while the water density from the simulations is shown in colored mesh. Non-displaceable water sites can be easily identified by visualizing the water occupancy from the simulations, such as in the case of site A, where the water occupancy directly overlaps with the three observed crystallographic water positions. ΔG_{SPAM} values from the water-only simulations classify only one of these three sites to be favorable (having a negative ΔG_{SPAM}), while the other two sites are unfavorable. However, none of the probes tested were capable of displacing these sites, which is consistent with their high experimental conservation.

Water sites displaced by one or more probes can also be easily identified by the lack of observed water density, indicated by the absence of colored mesh. For example, water site B is predicted to be displaced by isopropyl alcohol, *N*-methylacetamide, pyrimidine, and the acetate/methylammonium mix, but not by acetonitrile as water density in the presence of this probe molecule is observed. The MixMD predictions are consistent with experimentally observed displacement of this site by a carboxyl group in most ligands, as seen with the representative ligand in **Figure 4.3**[190]. This site is also known to be displaced by phosphate as in the structure of aldose reductase bound to glucose-6-phosphate (PDB:2ACQ), pyrrolidine of minalrestat (PDB:1PWL), and pyridazinone when bound to sulfonyl-pyridazinone inhibitors (PDB:1Z89)[191-193]. In some ligand-bound structures, the carboxyl group occupying site B is shifted to the right, and an additional water site C is observed[194]. This water site is found in the MixMD simulations of water and acetate/methylammonium, as shown in **Figure 4.3**. Site E is

predicted to be displaced by all of the tested probes, and is observed to be displaced by several functional groups, including thiazole derivatives (PDB:2NVC), chlorine of dicholorophenyl (PDB:2IPW), and by the sulfur substituent in epalrestat (PDB:4JIR)[190, 195, 196]. This is consistent with the MixMD predictions in which multiple functional groups are able to displace these water sites. In addition to known sites, MixMD identifies two hydration sites (D) which are not found in crystal structures of aldose reductase. These two sites are found within the water-only simulations to have positive (unfavorable) ΔG_{SPAM} values. This is consistent with the ligand-bound structures placing a carboxyl group at this site, but this interaction is not replicated by the acetate occupancy during the simulations. However, it is very possible that these sites are at least partially occupied by water molecules in solution, as crystallographic conditions such as temperature and resolution influence the number of identified water sites. For example, crystal structures solved at 2 Å resolution have ~1 water molecule per residue, while structures solved at 1 Å have ~1.6 waters per residue[197]. Therefore, it is likely that additional sites found in MixMD simulations represent hydration sites found in solution that are not identified in crystallography because of weak occupancy and experimental limitations.

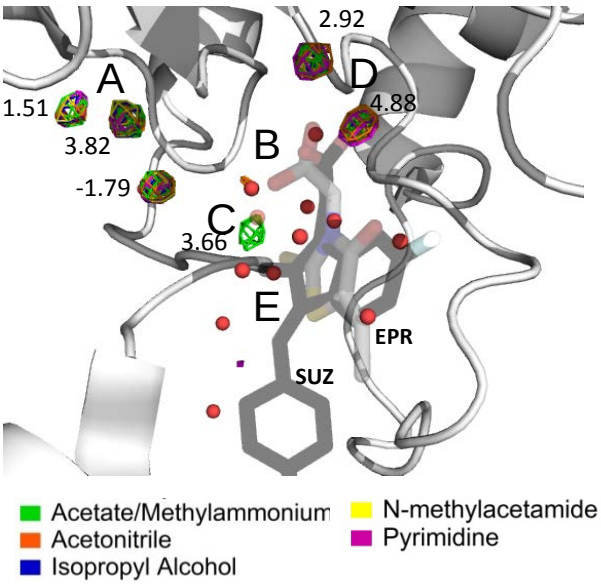


Figure 4.3: Aldose Reductase
Water density is shown at the 20 σ level, colored according to the probe type included in the simulation. ΔG_{SPAM} values from the water-only simulations are given in kcal/mol. ΔG_{SPAM} values for all waters within the active-site region ranged from -1.79 to 5.76 kcal/mol. Crystallographic waters (PDB:1ADS[172], 3U2C:WAT1338[194]) are shown for reference. Selected sites are labeled. The ligands epalrestat (PDB:4JIR,EPR[190]) and sulindac (PDB:3U2C,SUZ[194]) are shown for comparison. A) Cluster of water sites which are predicted by the MixMD simulations to be always conserved. B) Water site which is displaced by all probes except for acetonitrile. In some apo and ligand bound structures, a water molecule is found at site C, (PDB:3Q67:WAT710[198], 3U2C:WAT1338[194] transparent red sphere). When bound in this conformation, the oxygen of the ligand is positioned at site D. D) Water occupancy maxima not found in crystal structures. E) Water site displaced by all probe types in MixMD and ligands in crystal structures.

β-secretase

In the structure of β-secretase, the MixMD method correctly predicts several conserved water sites which are found in greater than 95% of crystal structures. The method also predicts the displacement of several water molecules known to be displaced in the majority of crystal structures. Interestingly, MixMD is able to predict the presence of a water molecule which bridges interactions between the ligand and protein in some cases, but is displaced by a ligand in others. In the apo-structure of BACE, this water molecule interacts with the two catalytic aspartates. As shown in **Figure 4.4A**, the amino group of inhibitors can displace this water site by mimicking this interaction, or this water site can be conserved to bridge interactions

between the protein and ligand, as shown in **Figure 4.4B**. The MixMD simulations predict that this water site can be displaced by the acetate/methylammonium, *N*-methylacetamide, and pyrimidine probes (as evidenced by the lack of water occupancy from these simulations). This is consistent with known ligands that use an amino group (PDB:4RCD, 3LL) to displace the water site by interacting with both aspartates[199]. Our simulations modeled the two aspartates as deprotonated, which has been shown to be the preferred protonation states for a subset of BACE inhibitors[200]. Other inhibitors, including those which place a hydroxy group at this site, preferentially interact with BACE when one of the aspartates is protonated[200, 201]. The MixMD results generated from the simulations with doubly deprotonated aspartates are consistent with this, which predicted this site to be favorably conserved in the presence of isopropyl alcohol. Using SPAM, the ΔG for this site is calculated to be -4.69 kcal/mol, which indicates that the site can be favorably occupied by a water molecule. This is consistent with the selective conservation of this site within ligand-bound structures. However, this site can also be displaced by some ligands, which is not apparent from the SPAM calculations alone. This water site has also been previously analyzed with the WaterMap method to guide synthesis efforts[202]. One of the goals of that study was to develop BACE inhibitors that did not displace the catalytic water in order to reduce the number of hydrogen bonds present in the inhibitor, in order to yield a ligand with more desirable drug-like properties. While the WaterMap method was successfully applied to explain SAR results, complementary structure-based drug design efforts were required. Alternatively, the MixMD method can be used, which allows users to predict the ease of displacement of a water site, while simultaneously predicting the location of favorable interactions of the probe molecules within the binding site. This information can then be used to identify favorable interactions that may be targeted with future ligands. Thus, the MixMD method yields additional information compared to other methods.

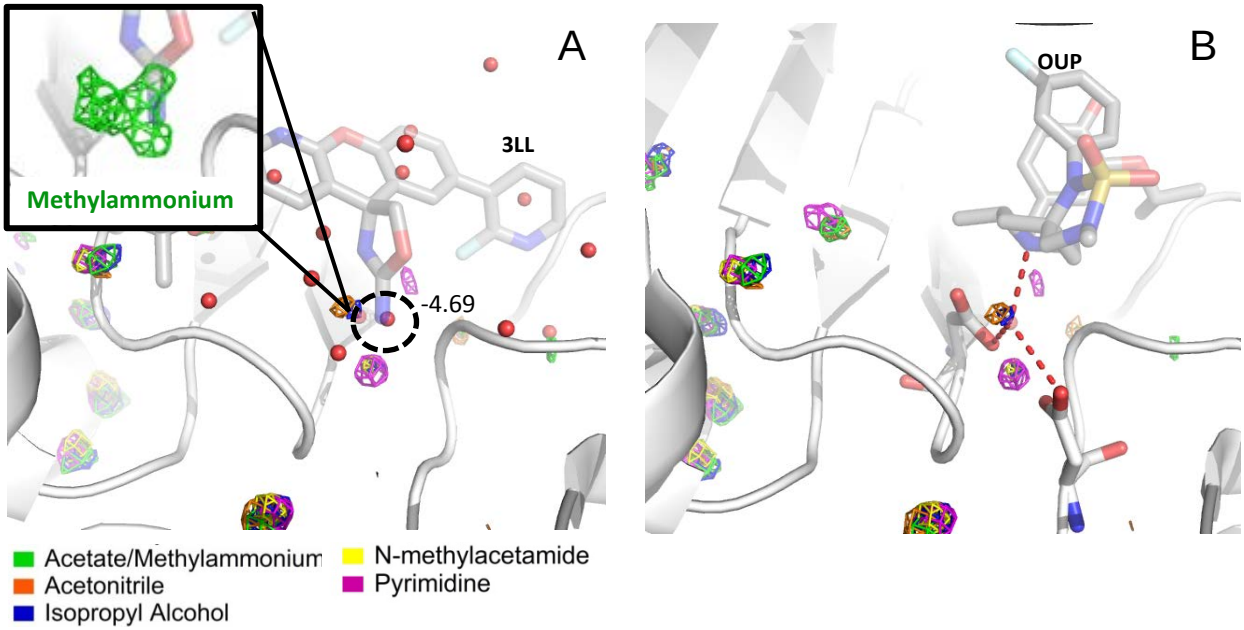


Figure 4.4: β-Secretase
Crystallographic waters from the apo structure of BACE (PDB: 1W50[174]) and the bridging water in the ligand bound structure (PDB:4FM7, WAT909[202]) are shown for reference. ΔG_{SPAM} for the circled water site is -4.69 kcal/mol in the water-only simulations. ΔG_{SPAM} values in the active site region ranged from -4.69 to 6.68 kcal/mol. A) MixMD correctly predicts the displacement of the circled water site by acetate/methylammonium, N-methylacetamide, and pyrimidine probes. The ligand from PDB:4RCD[199](3LL) is shown for comparison. Water density is shown at the 20 σ level, colored according to the probe type included in the simulation. The inset figure shows the Methylammonium density at 150 σ. B) This site may also be conserved and bridge interactions between the ligand and protein, as predicted by the simulations with acetonitrile and isopropyl alcohol. The ligand from PDB:4FM7[202](OUP) is shown for comparison.

β-lactamase

Apo β-lactamase contains a number of water sites which are experimentally known to be displaced in ligand-bound structures. These sites are correctly predicted by MixMD to be displaced by probes. In addition to the displaceable water sites in β-lactamase, the MixMD simulations also predict the location of conserved waters, including the cluster of water molecules known to be important in stabilizing the Ω-loop[203]. However, there are two exceptions, as shown in **Figure 4.5**. Classic inhibitors of β-lactamase form a covalent attachment to the enzyme following nucleophilic attack of the β-lactam ring by a deprotonated serine. The carbonyl oxygen of the β-lactam ring displaces a water molecule, while a nearby water molecule coordinated to Glu-166 is involved in hydrolysis of the β-lactam ring[204]. Both

sites are known to be displaced by boronic acid inhibitors[205]. SPAM calculations for the water-only simulations predict one of the sites to be very favorably occupied (ΔG_{SPAM}=-5.58) and the other to be unfavorable (ΔG_{SPAM} = 1.17). In contrast, MixMD simulations predict both of these sites to be conserved. This discrepancy is likely due to two reasons. First, simulations were performed with the serine protonated, which therefore mimics the apo structure in which the serine is free to coordinate with the water molecule rather than being covalently attached to the inhibitor. Second, the boronic acid inhibitors were designed as analogues of the transition state, and they are able to uniquely emulate the interactions of the hydration sites allowing for their displacement. These specific interactions are not replicated by the current probes in our methodology, and so these sites were predicted to be conserved. As the current set of probes is limited, future introduction of additional probes is expected to eliminate this shortcoming.

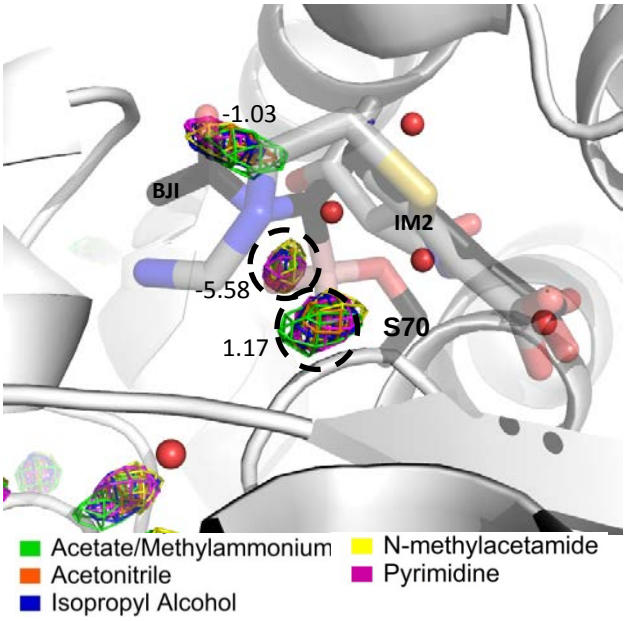


Figure 4.5: β-lactamase
Water density is shown at the 20 σ level, colored according to the probe type included in the simulation. Crystallographic waters from the apo structure (PDB:1ZG4[173]) are shown for reference. ΔG_{SPAM} values from the water-only simulations are shown in units of kcal/mol. ΔG_{SPAM} values for the active-site region ranged from -5.58 to 3.62 kcal/mol. While MixMD correctly predicts many of the waters in the active site of β-lactamase as being displaced, there are two known discrepancies. These are attributed to the limited set of probe types used and the inability to account for covalent interactions within an MD simulation. (PDB:1BT5-IM2[204], 1ERM-BJI[205])

BRD4

Apo BRD4 contains several water molecules which are displaced upon interaction with ligands. For example, site A in **Figure 4.6** is predicted to be displaced by all of the probe types tested. This is consistent with crystal structures of bound ligands, in which many functional groups displace this site, including triazole (PDB:2YEL, WSH), the carbonyl of acetylated histone proteins (PDB:3JVK, peptide), and the oxygen of isoxazole (PDB:3SVF, WDR)[206-208]. Within the binding pocket, there are a number of water molecules which are found in the majority of crystal structures. For example, site B in **Figure 4.6** is found in 97% of all comparable crystal structures. Interestingly, MixMD predicts that several of these sites can be selectively displaced. SPAM ΔG values for these sites are also positive, indicating their potential for displacement. A recent crystal structure has been solved verifying this, in which an inhibitor extends deeper into the binding pocket and displaces these sites (PDB:4O7F, 2RQ[209]), as shown in **Figure 4.6**. This emphasizes the utility of and need for water prediction methods. Based on conservation alone, it would appear that these sites are not easily displaced. However, MixMD simulations predict that they can be selectively displaced, in agreement with experimental data.

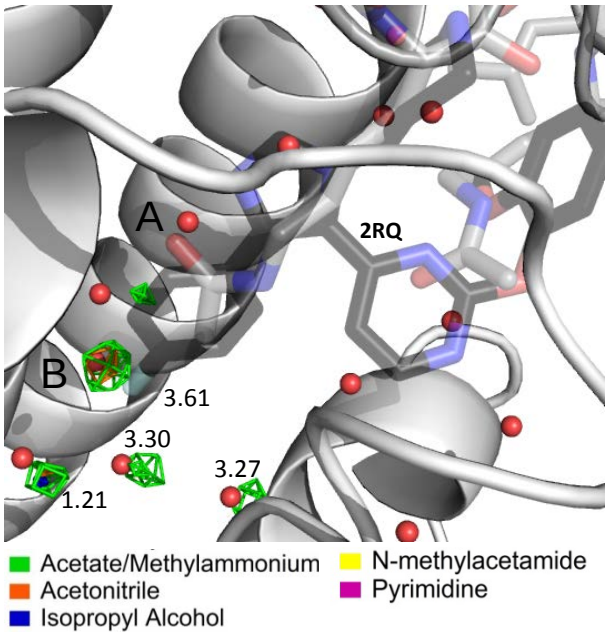


Figure 4.6: BRD4
Water density is shown at the 20 σ level, colored according to the probe type included in the simulation. Crystallographic waters from the apo structure (PDB:2OSS[175]) are shown for reference. ΔG_{SPAM} values from the water-only simulations are shown in units of kcal/mol. ΔG_{SPAM} values for the active site region ranged from -0.73 to 3.61 kcal/mol. A) Site predicted by MixMD to be displaced, shown with an example ligand (PDB:3UVW, peptide[175]) displacing the site. B) Water site found in 97% of comparable structures, predicted by MixMD to be displaceable is shown with an inhibitor displacing this site (PDB:4O7F, 2RQ[209]).

DHFR

The results for DHFR provide another good example of MixMD's ability to discriminate between waters that are always conserved, always displaced, and those that are selectively displaced. In DHFR, 100% of homologous structures contain a water at site A (**Figure 4.7**) which is predicted by MixMD to be always conserved. On the other hand, SPAM calculations predict this site to be unfavorable, with a ΔG_{SPAM} of 5.13 kcal/mol. Site B is often displaced, with only 18% of structures containing a water molecule at this location. This is consistent with SPAM calculations, which also identify this site as unfavorable (ΔG_{SPAM} of 5.13 kcal/mol). For example, the amino group of methotrexate (PDB:1DF7, MTX) displaces this site, which in agreement with the MixMD prediction that this site will be displaced by *N*-methylacetamide[176]. The inability of other groups to displace this site is illustrated in the binding mode of folic acid to DHFR. Folic acid is almost identical in composition to methotrexate, with differing substituents at two sites,

but binds with a different orientation[210]. In methotrexate, a nitrogen (which occupies site B on the crystal structure) substitutes for an oxygen in folic acid. However, folic acid binds to DHFR with the pteridine ring flipped 180°, which results in the oxygen pointing in the opposite direction. This specificity is captured by the behavior of the probes in the simulations. Visualizing the *N*-methylacetamide occupancy by atom shows that the nitrogen is oriented in the direction known to be preferred from ligand-bound structures with the oxygen always positioned away from this site. Site C is another example of nitrogen displacing a water molecule. In this case, MixMD predicts that this site can also be displaced by *N*-methylacetamide, as well as acetate/methylammonium and isopropyl alcohol. Thus, not only can MixMD identify displaceable water sites, but can also identify specific functional groups capable of displacing a site.

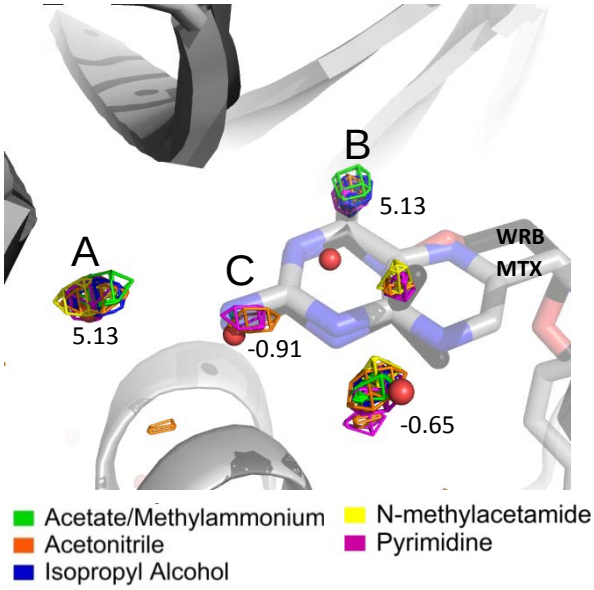


Figure 4.7: Dihydrofolate Reductase
Water density is shown at the 20 σ level, colored according to the probe type included in the simulation. Crystallographic waters from the apo structure (PDB:1DG8[176]) are shown for reference. ΔG_{SPAM} values from the water-only simulations are shown in kcal/mol. ΔG_{SPAM} values within the active-site region ranged from -1.42 to 5.13 kcal/mol. A) Water that is found in 100% of comparable crystal structures, predicted to be conserved by MixMD. B) Water site known to be displaced by nitrogen, predicted by MixMD to be displaced by N-methylacetamide. C) Water site known to be displaced by nitrogen, predicted by MixMD to be displaced by N-methylacetamide, acetate/methylammonium, and isopropyl alcohol. (PDB:1DF7[176](MTX) and 1DG7[176](WRB))

HSP90

HSP90 has been well studied, and many potent inhibitors exist[211]. Site A, as shown in **Figure 4.8**, is found in 100% of homologous structures and is predicted by MixMD simulations to be favorably conserved. SPAM calculations identify this site as being only weakly favorable (ΔG_{SPAM} of -0.34 kcal/mol). Studies focused on the structure-activity relationship of HSP90 have noted the tightly coordinated nature of this water molecule, leading researchers to avoid displacing this site[212]. Site B, on the other hand, is displaced by ligands containing either a hydroxyl group or a carbonyl group, as shown with geldanamycin bound to HSP90 in **Figure 4.8**[213]. This is consistent with the MixMD prediction that this site is displaceable by N-methyl acetamide, with the hydrogen-bond donor and acceptor regions of *N*-methylacetamide occupying similar orientations to those found in ligand bound structures[177, 214]. HSP90 has previously been studied by Alvarez-Garcia and Barril using their cosolvent simulation method MDmix[99], and by Haider and Huggins using IFST with MCSS[98]. IFST with MCSS incorrectly predicted site B to be conserved based on predicted ΔG values. SPAM calculations also identify this site as being weakly favorable (ΔG_{SPAM} of -1.04 kcal/mol), whereas MDmix correctly predicts site B as displaceable (1AH6:393, 1YER:336). In Barril's MDmix, a water site is classified as displaceable if one of the tested probe molecules binds to the site with higher affinity. However, in MDmix only ethanol and acetamide solvent mixtures were used, which limits the applicability of the data. For example, water 391 (PDB:1AH6) in the crystal structure of HSP90 is displaced by the phosphate groups of ATP (PDB:1BYQ)[215]. Barril's MDmix predicts this water to be conserved (water site 325 in 1YER numbering), as none of their probes are capable of displacing this site, while our method correctly predicts this site as displaceable. Thus, the use of multiple probe molecules in MixMD offers a greater predictive power over alternative methods that utilize a more limited set of probes.

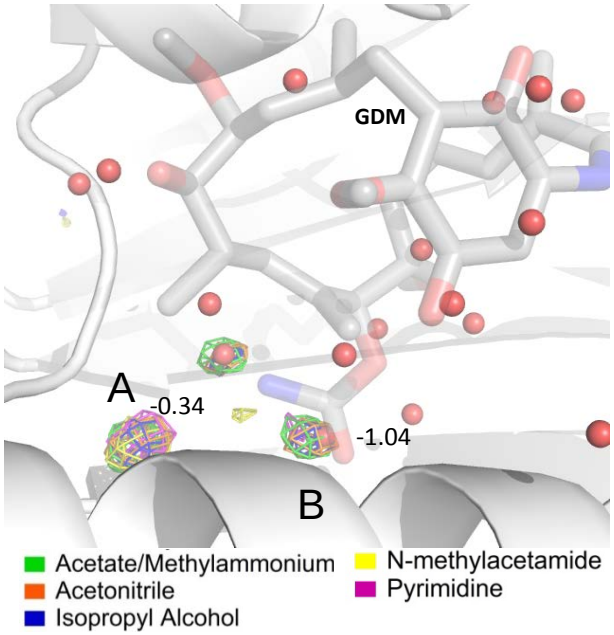


Figure 4.8: Heat Shock Protein 90
Water density is shown at the 20 σ level, colored according to the probe type included in the simulation. Crystallographic waters (PDB: 1AH6[177]) are shown for reference. ΔG_{SPAM} values from the water-only simulations are shown in kcal/mol. ΔG_{SPAM} values within the active-site region ranged from -1.65 to 5.70. Geldanamycin (PDB:2YGA,GDM[213]) is shown for reference A) Water site found in 100% of homologous structures, predicted to be conserved by MixMD. B) Water site displaced by carbonyl of geldanamycin, predicted to be displaced by N-methylacetamide.

Neuraminidase

Upon ligand binding to neuraminidase, several water sites are conserved. For example, a cluster of water molecules, shown in **Figure 4.9A**, are found in 100% of homologous crystal structures. SPAM calculations predict that these sites are unfavorable, with ΔG_{SPAM} values ranging from 0.48 to 8.56 kcal/mol. However, MixMD simulations predict the conservation of these sites in the presence of all probes tested, consistent with their high experimental conservation. On the other hand, a number of waters are displaced upon ligand binding, for instance by the carboxyl group of the ligand, as shown in **Figure 4.9B**[178]. MixMD correctly predicts the displacement of these sites, as indicated by the lack of water density at this location and the presence of acetate density. Although other methods have been applied to neuraminidase, they require additional steps to generate comparable information. For

example, neuraminidase has been previously studied by the JAWS method[87]. While the JAWS method was able to identify favorable and unfavorable hydration sites in the active site, the method requires the use of ligand-bound structures to identify water sites that would be displaced upon ligand binding. Our MixMD method does not require ligand-bound structures, and all of these simulations were initiated from apo structures. MixMD simulations could be easily extended to study sequence level changes. For example, neuraminidase variants are common, and show differing susceptibilities to inhibitors[216]. Interestingly, the number of water sites contained in the active site has been shown to vary depending on the mutant studied, and it has been suggested as one factor influencing the observed variations in binding affinity of inhibitors[217]. MixMD could potentially be used for further study of neuraminidase variants, to yield insight into the specific factors that mediate the observed water occupancy and variable binding affinities of inhibitors.

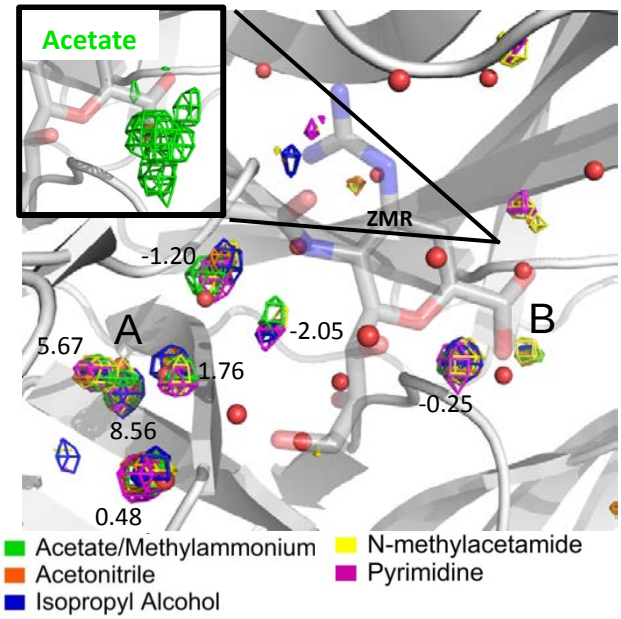


Figure 4.9: Neuraminidase
Crystallographic waters (PDB:4HZV[178]) within 10 Å of the MixMD-identified hotspot are shown, along with water density from the MixMD simulations shown at the 20 σ level, colored according to the probe type included in the simulation. ΔG_{SPAM} values from the water-only simulations are shown in kcal/mol. ΔG_{SPAM} values within the active-site region ranged from -3.35 to 8.56 kcal/mol. A) Cluster of conserved water sites found in 100% of homologous structures, predicted by MixMD to be conserved. B) Water sites displaced by carboxyl of ligand (Zanamivir shown, PDB:4I00, ZMR[178]) are correctly predicted by MixMD to be displaced. The inset figure shows the occupancy of the acetate probe which correctly predicts displacement of these sites.

Penicillin Binding Protein

The MixMD results for PBP-4 provide another example of the method's ability to predict conserved and displaceable sites, as well as to discriminate between related systems. As shown in **Figure 4.10**, the MixMD results correctly predict the conservation of the water at site A, which is found in all related structures. The results also correctly predict the displacement of water at site B, which is displaced by the carbonyl oxygen of the β-lactam ring of penicillin derivatives[179]. SPAM calculations identify site B as weakly favorable (ΔG_{SPAM} of -0.86 kcal/mol) while site A is identified as unfavorable (ΔG_{SPAM} of 9.55 kcal/mol). Interestingly, TEM-1 β-lactamase has an equivalent water molecule to site B which was predicted to be conserved. In TEM-1, this water molecule interacts with the nearby hydrolytic water molecule, the backbone carbonyl of Ala237, and the backbone nitrogens of Ser70 and Ala237[173]. The surrounding environment of PBP-4 is similar, although the potential interactions with the neighboring water molecule are lost since PBP-4 does not contain an analogous water molecule at this site[179]. The difference in observed occupancy values can therefore be rationalized as being due to differing coordination of the two water molecules, consistent with the apo crystal structures of TEM-1 and PBP-4. While TEM-1 and PBP-4 both interact with β-lactam rings, they have evolved to have different functions[218]. The fact that the MixMD results for these two enzymes are not identical illustrates the value and usefulness of MixMD to be applied in the design of specific inhibitors.

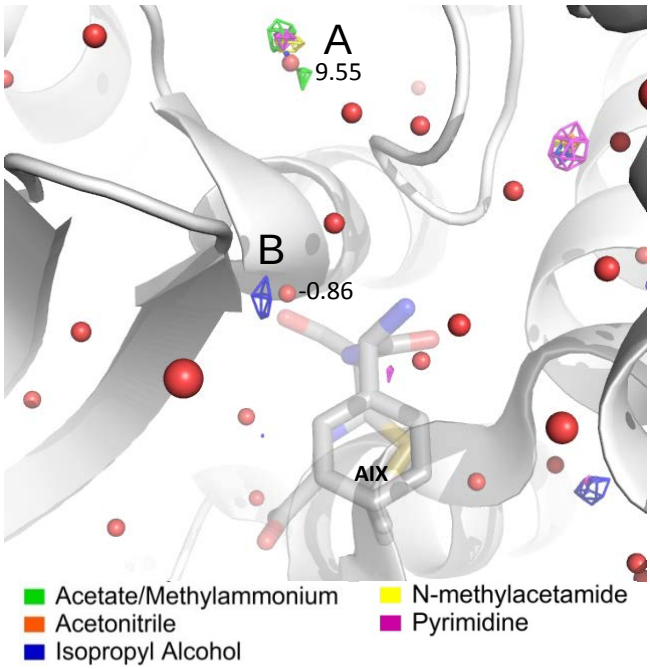


Figure 4.10: Penicillin Binding Protein

Crystallographic waters (PDB:2EX2[179]) within 10 Å of the MixMD identified hotspot are shown, along with water density from the MixMD simulations shown at the 20 σ level, colored according to the probe type included in the simulation. ΔG_{SPAM} values from the water-only simulations are shown in kcal/mol. ΔG_{SPAM} values within the active-site region ranged from -1.09 to 9.55 kcal/mol. A) Water site found in 100% of related crystal structures, predicted to be conserved in the presence of all probe types tested. B) Water site displaced by ligand (PDB:2EX6, AIX shown[179]) is predicted to be displaced by all probes other than isopropyl alcohol.

Penicillopepsin

Ligands of penicillopepsin displace a number of water sites, as shown in **Figure 4.11**. Within the active-site region, only two water sites are predicted as being conserved, while all other sites are predicted to be potentially displaceable. Aspartic proteases, such as penicillopepsin, unvaryingly have a water molecule that interacts with the two active aspartates and is involved in catalysis[219]. This location is shown at site A in **Figure 4.11**. However, this water may be displaced by inhibitors that interact with these aspartates. For instance, this water is displaced by the phosphonate portion of the ligand shown in **Figure 4.11**[220]. This is consistent with the MixMD predictions that this site will be displaced. As MixMD incorporates both charged and uncharged probe molecules, the method is able to predict the displacement

of water sites that commonly bind charged ligands, as shown in the case of site A. Additionally, MixMD predicts the displacement of several other water sites, consistent with ligand-bound crystal structures which show the majority of water sites in this region to be displaced upon binding. MixMD also predicts the location of water sites which are known to be conserved, including the water located at site B. On the other hand, SPAM calculations predict this site to be unfavorable, with a ΔG_{SPAM} of 7.71 kcal/mol. This water molecule is buried and participates in a network of interactions that are essential to stabilize the active site[219]. It is conserved in 100% of related structures of penicillopepsin as well as in structures of related aspartic proteases. Furthermore, disruption of this stabilizing network of interactions has been show to disrupt the active-site geometry in related enzymes[219], illustrating the biological importance in conserving this water site.

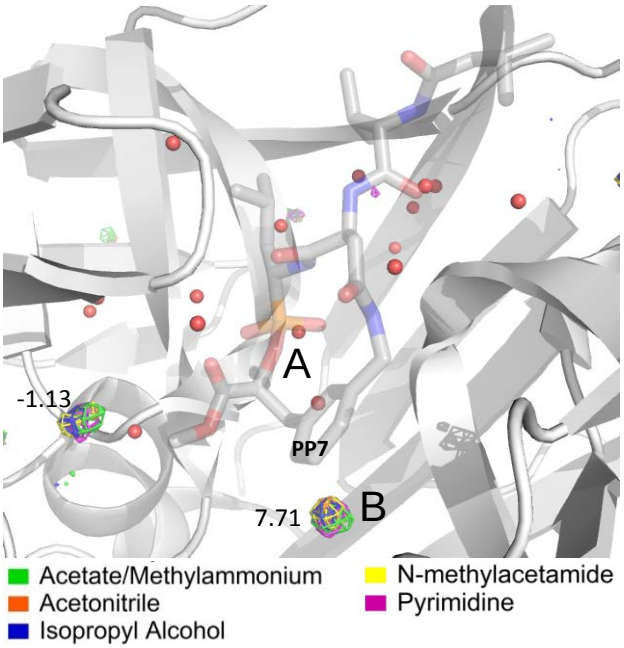


Figure 4.11: Penicillopepsin
Crystallographic waters (PDB:3APP[180]) within the active site are shown, along with the water density from the MixMD simulations at the 20 σ level, colored according to the probe type included in the simulation. ΔG_{SPAM} values from the water-only simulations are shown in kcal/mol. ΔG_{SPAM} values within the active-site region ranged from -1.17 to 7.71 kcal/mol. A) Water site displaced by phosphonate-containing ligand (PDB:1BXO, PP7[220]) is correctly predicted as displaceable by the MixMD simulations. B) Important water site found in 100% of related structures which participates in a network of stabilizing interactions is predicted as being conserved.

Thrombin

Upon ligand binding to thrombin, several water sites are displaced, as shown in **Figure 4.12**[181]. The majority of water sites within this region have positive ΔG_{SPAM} values, consistent with their favorable displacement. Interestingly, a number of water sites are observed in the MixMD simulations which are known to be involved in thrombin's activity. Thrombin is allosterically regulated by a Sodium ion, whose binding site is connected to the active site via a water channel[221, 222]. Site B is located within this region, and it is identified by the MixMD simulations to be conserved, although SPAM calculations classify this site as unfavorable (ΔG_{SPAM} of 1.92 kcal/mol). One of the benefits of the MixMD methodology is the ability to contour occupancy at different levels, corresponding to a range of very high to moderate to low occupancy. While high σ levels in the presence of probe molecules were used as a cutoff for classifying water conservation, lower σ values still identify discrete water sites with occupancy greater than that of bulk water. When the water occupancy is visualized at lower occupancy levels, such as 10 σ, several additional sites within the water channel of thrombin are identified, pointing to MixMD's ability to identify not only absolutely conserved sites, but also water sites that will be occupied in the absence of bound ligands.

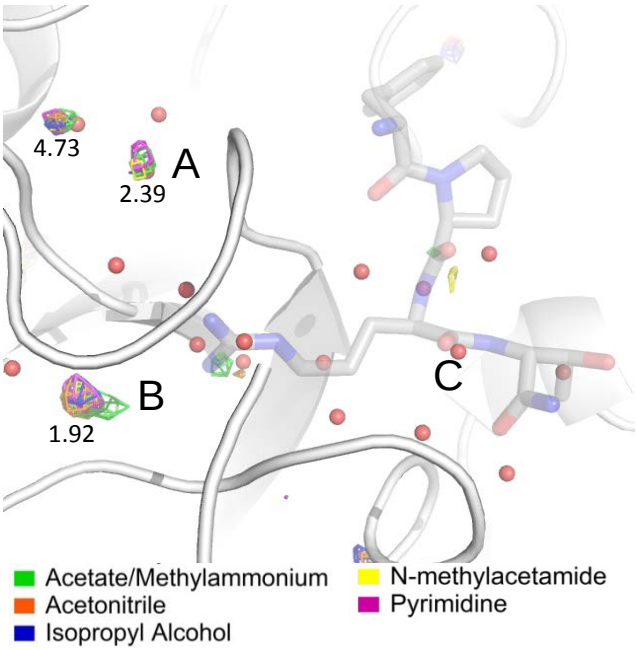


Figure 4.12: Thrombin
Crystallographic waters within the active site (PDB:3U69[181]) are shown, along with the water density from the MixMD simulations at the 20 σ level, colored according to the probe type included in the simulation. ΔG_{SPAM} values from the water-only simulations are shown in kcal/mol. ΔG_{SPAM} values within the active-site region ranged from -0.33 to 5.63 kcal/mol. A) Water site that is found in 74% of comparable crystal structures and is predicted to be selectively displaced by acetonitrile and isopropyl alcohol. B) Water site is predicted to be always conserved, found in 100% of comparable crystal structures. C) Water site that is predicted to be always displaced, shown with a peptide-inhibitor. (PDB:3U8O[181]).

4.5 Conclusions

As shown in the examples above, the MixMD method correctly predicts the conservation and displacement of the water sites in each system tested. We have also shown that the SPAM method applied to water-only simulations yields affinities inconsistent with experimental data. This shows that a water molecule's binding affinity alone cannot account for the ability of a particular functional group to replicate the specific interactions coordinating each water molecule. Although ligands may be designed to displace a water site, this is not necessarily accompanied by a corresponding increase in binding affinity if the ligand does not adequately mimic the specific contacts previously made by the water molecule. Using the MixMD method, favorable binding sites on the protein's surface are determined for multiple

functional groups. This in turn allows for the prediction of conserved and displaceable water sites, while simultaneously determining which groups can successfully displace them. MixMD is able to identify specific groups that can displace a site, identify conserved water sites that play important roles and are involved in protein function, and discriminate between closely related proteins, including β-lactamase and penicillin binding protein. In addition, MixMD correctly predicts the displaceability of water sites that are incorrectly predicted by other methods as being conserved, as shown in the results of HSP90. MixMD had only one shortcoming in the β-lactamase case. Boronic acid inhibitors displace two water sites in β-lactamase which were predicted as conserved in the MixMD simulations. None of the probes in the current set contain diols, and so the displacement of this site was not predicted. Efforts are currently underway to expand the available probe set to include additional groups, which is expected to extend MixMD's predictive power. Overall, the MixMD method successfully classifies the displacement of water sites by common functional groups. These results may be used in the strategic design of ligands to determine which water sites should be conserved and which sites can be favorably displaced. Furthermore, MixMD results can also give insight into pockets that ligands may be most favorably extended into, by predicting sites that are favorably desolvated.

Chapter 5. MixMD Pharmacophore Development and Application

5.1 Abstract

Receptor-based pharmacophore models describe the location and extent of favorable interaction sites on a protein's surface. Pharmacophore models are generated by examining the interactions made by individual functional groups, either from mapping methods based on static structures or using molecular dynamics simulations in the presence of small molecule probes. Our group has developed the mixed-solvent molecular dynamics (MixMD) method for identifying specific favorable interactions and binding-site regions on a protein's surface. MixMD and related cosolvent methods are an especially promising means of mapping binding sites, as they explicitly account for the role that protein dynamics and solvent play in mediating protein-ligand interactions. We have developed a framework for converting the occupancies of specific functional groups on the cosolvents into pharmacophore features. These pharmacophore features are then consolidated into pharmacophore models for use with the program MOE. The pharmacophore models can then be screened against libraries of compounds to identify potential new inhibitors that replicate the desired interactions. Using ABL kinase as a test system, we show a good ability to discriminate between active and inactive compounds based on MixMD generated pharmacophore models. In every pharmacophore model tested, a larger proportion of active compounds satisfied the model than did inactive compounds. Prospective application of this method to develop allosteric ligands for Src kinase is currently underway.

5.2 Introduction

Pharmacophore models map the important features responsible for the interaction between a ligand and its target. Pharmacophore models can be generated from related ligands, termed ligand-based pharmacophores, to describe conserved features within the series of ligands. Models can also be generated in relation to the protein's surface, which are known as receptor-based pharmacophore models.

Several methods have been proposed to identify important interactions on a protein target. The earliest methods focused on the interaction between small molecule fragments ("probes") and a protein's surface. The aptly named GRID method utilizes a 3-dimensional grid to describe the space surrounding the protein surface. At each grid point, the potential energy of the probe molecule is calculated, thereby identifying favorable interaction sites[47]. Similarly, MCSS (Multiple Copy Simultaneous Search) performs minimization of a few thousand small molecule probes which are initially scattered across the active site. The resulting minima have been shown to overlap with functional groups of known ligands[48]. More recently, FTMap has been introduced, which also focuses on the identification of minima of small molecule probes[49, 223]. FTMap is implemented as a webserver where users upload their target of interest. Potential interactions with the target are sampled via rigid body docking of 16 probe molecule types. The resulting locations are then ranked based on their calculated energy. While these methods are very fast, they are not able to identify all known ligand binding sites, likely because of insufficiently accounting for protein flexibility[51].

Cosolvent molecular dynamics (MD) simulations are a promising means of generating receptor-based pharmacophore models, as they explicitly account for the effects of solvation and protein flexibility. Cosolvent MD methods use a mix of small molecule probes and water to solvate the protein of interest. MD simulations allow for sampling of potential probe and water positions while simultaneously accounting for conformational flexibility of the protein. Following molecular dynamics, the occupancy of probe molecules over the course of the

simulation is determined. Sites on the protein's surface that are frequently occupied by multiple probe molecules correspond to regions making important interactions, and therefore identify biologically relevant sites. Although several cosolvent methods have been introduced, as recently reviewed by Ghanakota and Carlson[62], few have actually been developed enough for successful pharmacophore modeling.

The SILCS method by Mackerell and coworkers has had the most comprehensive development. SILCS uses fairly high concentrations of small molecule probes to map protein surfaces. Using either traditional MD or a combined Monte-Carlo/MD approach, benzene, propane, methanol, formamide, acetaldehyde, methyl-ammonium and acetate probes at a concentration of ~0.25 M each are allowed to sample potential binding sites on the protein's surface[65, 67]. The resulting trajectories are aligned and overlaid with a 1 Å cubic grid, and the occupancy of the probe molecules at each grid point is determined. In the SILCS method, occupancies are converted into free energy values at each grid point based on the Boltzmann distribution[64]. Grid points having favorable (negative) energy values within the region of interest are selected for construction of the pharmacophore model[71]. The selected grid points are then clustered using a distance-based algorithm, to define distinct interaction sites. Pharmacophore features are created from these sites using a sphere centered on the average of the grid point locations, with a radius set to contain all of the grid points within the cluster up to a maximum value. This upper limit for pharmacophore radii prevents excessively large pharmacophore features, which would hinder specificity of the resulting model. Following pharmacophore feature generation, potential pharmacophore models are created by joining the pharmacophore features into several groups. These models are ranked by the number of pharmacophore features they contain and by the sum of the grid free energies of each of the pharmacophore features. The SILCS pharmacophore method was initially validated on HIV protease, Factor Xa, and dihydrofolate reductase. For the three systems, known ligands were preferentially selected over decoy ligands, with AUCs ranging from 0.56 to 0.88[71]. Subsequent SILCS pharmacophore validation studies using additional systems and a larger number of

solvents showed similar performance[70]. While the SILCS pharmacophore method showed reasonably good performance, it is a commercial service and not freely available.

We have focused on the extension of the MixMD method developed by our group for the generation of receptor-based pharmacophore models. The MixMD method has been carefully developed to map known binding sites on a protein's surface while accounting for protein flexibility[51, 52]. One of the important differences in the MixMD method relative to others is the use of fairly low (5% by volume) concentrations of water-miscible, organic solvent probe molecules. Using low concentrations of probe molecules has been shown to accurately reproduce known binding sites, while decreasing the number of false-positive sites identified[79]. Importantly, using low concentrations of organic solvents also allows for experimental validation, as most crystallographic studies would not be feasible at higher concentrations. Previous MixMD studies have introduced several potential solvents, including imidazole, pyrimidine, acetonitrile, isopropyl alcohol, n-methylacetamide, methyl-ammonium, and acetate[51, 73]. These solvents give insight into a wide range of potential interactions taking place on the protein's surface, including hydrogen-bonding, hydrophobic, aromatic, and charged interactions. In order to validate the MixMD pharmacophore method, ABL kinase was selected as a test system. Successful validation of the method ensures that virtual screening performed using MixMD generated pharmacophore models will be capable of distinguishing true ligands from inactive compounds. Following validation, the MixMD method was prospectively applied to Src kinase towards the development of novel allosteric inhibitors.

ABL and Src share a similar overall structure, containing kinase, Src homology 2 (SH-2) and Src homology 3 (SH-3) domains. Most ligands bind within the active-site region of the kinase domain, although a few allosteric ligands have been developed. For example, in the inactive form of ABL the c-terminus of the kinase domain adopts a helical structure and folds back against a myristate bound pocket nearby. This conformation allows the SH-2 and SH-3 domains to close against the kinase domain. Ligands designed to replicate this interaction stabilize the closed form, thereby acting as allosteric inhibitors[224]. This interaction is absent in

Src, which instead has a phosphorylated tyrosine that serves as a "latch" for the SH-2 domain (**Figure 5.1**). With the goal of identifying similar acting sites in Src that would also promote the closed form, this form of Src was used as input for MixMD simulations.

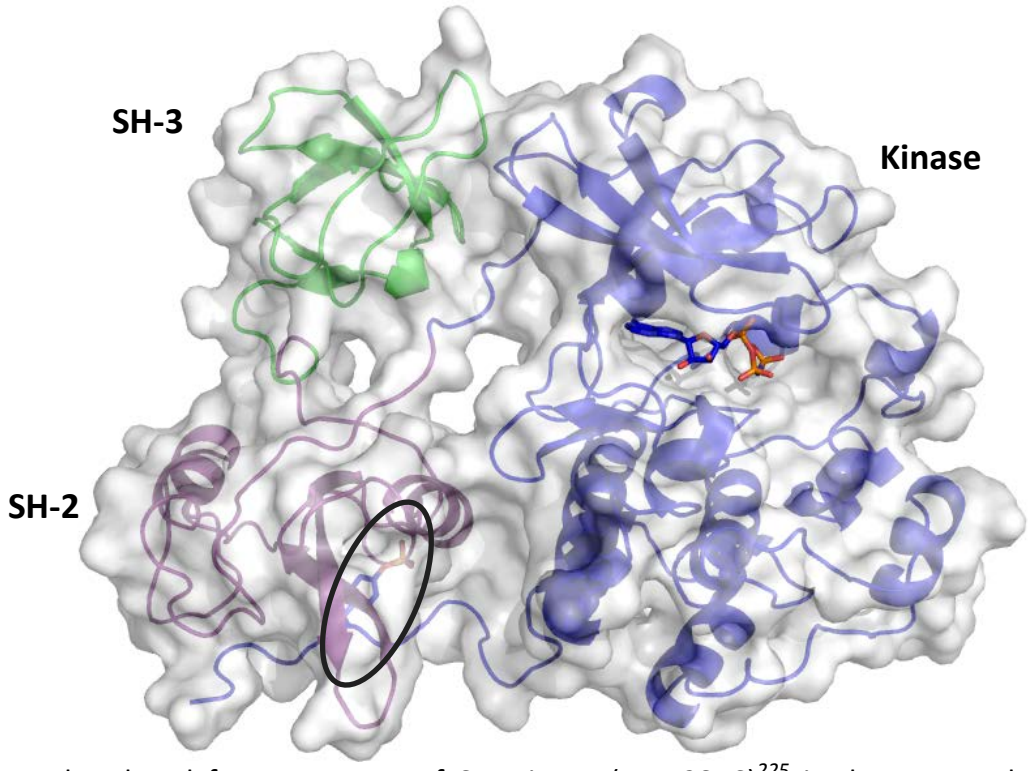


Figure 5.1: The closed form structure of Src Kinase (PDB:2SRC)[225] is shown. In the closed conformation, a phosphorylated tyrosine (circled) at the very c-terminus of the kinase domain binds to the SH-2 domain. In the open form, this interaction is absent and the SH-2 and SH-3 domains rotate away from the kinase domain. Most kinase inhibitors target the ATP-binding site within the kinase domain.

5.3 Methods

Simulation procedure

Simulations of ABL kinase have been previously completed by our group using pyrimidine, isopropyl alcohol, and acetonitrile[51]. For completeness, additional simulations using imidazole, n-methylacetamide, and a methylammonium/acetate mixture were completed for ABL to incorporate all MixMD solvent types into the MixMD pharmacophore method. System setup was as previously described[51]. Simulations of SRC Kinase were initiated from a

crystal structure of the closed form containing the SH2, SH3, and kinase domains (PDB:2SRC)[225]. In the closed form, tyrosine-527 near the c-terminus is phosphorylated and binds to the SH2 domain. This residue was phosphorylated in the simulations. The crystal structure contained an ATP analog, AMP-PNP, which was removed. Hydrogens were added and side chain positions were optimized using MolProbity[182]. Simulations were prepared using four solvent mixtures: pyrimidine, acetonitrile, isopropyl alcohol, and an acetate/methylammonium mixture[73]. These five solvent types were chosen to identify aromatic, hydrophobic, hydrogen-bonding, and charged interactions. Each solvent mixture was run individually, and was initiated as a layer of solvent probe molecules surrounding the protein followed by a box of TIP3P water in a 5%/95% ratio of solvent to water. Sodium ions were added to yield a net neutral charge. Structures were prepared in the AMBER utility tleap, using the ff99SB force field[185]. For each solvent mixture, ten individual simulations were performed. Minimization was done for 5000 steps with restraints on the protein, followed by 2500 steps of minimization on the entire system. Following minimization, the systems were heated to 300K at constant volume over 40,000 steps with a 2 fs timestep and restraints of 10 kcal/mol-Å^2 on the protein. The systems were then equilibrated at constant pressure for 1.75 ns as the restraints were gradually removed. Ten production runs for each probe mixture were performed for 20ns using the GPU enabled version of AMBER11/12 PMEMD[185, 226-228].

Identifying Binding Sites

MixMD simulations have previously been shown to identify biologically relevant sites as those having high levels of probe occupancy from multiple probe types[51]. Active and allosteric sites typically fall within the top few (~5) ranked sites. In order to determine the probe occupancy, the last 10 ns of each trajectory were aligned with the cpptraj utility in AMBER[157]. Using a 0.5 Å x 0.5 Å x 0.5 Å grid, the occupancies of each probe molecule at every point on the protein's surface were counted. The results were then normalized into units of standard deviations away from the mean occupancy (termed σ units). Occupancy maps were visualized in PyMOL[100] to identify regions on the protein's surface having high occupancy arising from multiple, overlapping probe types.

Pharmacophore Development

Once potential binding sites are identified, the interactions within each binding site region can be converted into pharmacophore models for use in virtual screening. To do so, the resulting trajectories are aligned, the protein is overlaid with a 0.5 Å cubic grid, and the occupancies for each grid point are determined (**Figure 5.2**). Because pharmacophore models require specific interaction types rather than overall probe locations, the occupancy for pharmacophore models was calculated for individual functional groups rather than total probe occupancy. For pyrimidine, imidazole, and acetonitrile, the occupancy was determined from the center of mass of each of the molecules. Counting the occupancy in this manner yields the center of the aromatic or hydrophobic interaction site. The oxygen and methyl groups of isopropyl alcohol and oxygen and nitrogen of N-methylacetamide were each counted separately to account for the presence of multiple functional groups within each probe. Acetate and methylammonium probe occupancy was determined based on the location of all atoms, to identify regions favoring charged interactions. Highly charged ligands are typically undesirable, so these features were not included in the final pharmacophore models used for virtual screening. However, they were incorporated in the MixMD pharmacophore generation protocol, and so may be incorporated into pharmacophore models when desired.

Prior to converting the grid points into pharmacophore features, low occupancy grid points were removed. This facilitates identification of high affinity interactions, corresponding to high occupancy regions. Including the low occupancy points would result in extremely large pharmacophore features that encompass the majority of the protein's surface. The cutoff for each set of grid points (in the range of 10-20% of the maximum value) was chosen based on visual inspection, so that discrete sites were obtained rather than large patches of density. For ABL kinase, a cutoff of 10% was used for all maps except for acetate, methylammonium, and the n-methylacetamide carbonyl maps which used points having greater than 15% occupancy. For Src Kinase, acetonitrile, pyrimidine, and isopropyl alcohol maps were cutoff at 15% of the maximum value while 20% was chosen as the cutoff for acetate and methylammonium

occupancy. An in-house R script was created to group grid points into clusters using the DBSCAN algorithm[229]. DBSCAN is a density based clustering algorithm that identifies connected regions of density within a set distance parameter, eps. This was set to 1 Å in the current protocol. The process of clustering individual probe occupancy is depicted in **Figure 5.2**.

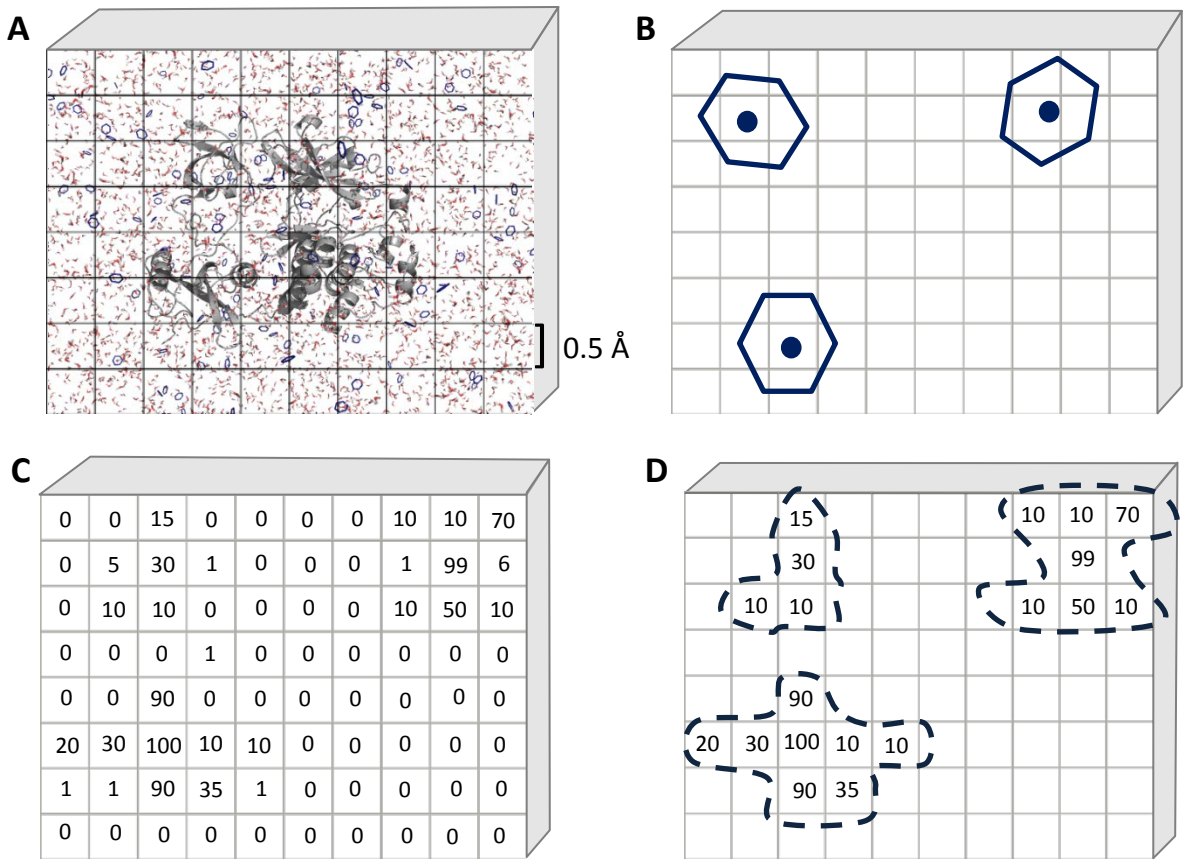


Figure 5.2: A) Individual trajectories are aligned and overlaid with a 0.5 Å cubic grid. B) At each grid point, the occupancy of probe molecules is counted for each frame in the trajectory. For example, the occupancy of the center of aromatic probes is counted at each grid point. C) This yields a time-averaged occupancy value at each grid point. D) Low occupancy grid points are removed (eg. those less than 10% of the max occupancy). The remaining points are clustered with the DBSCAN algorithm to identify discrete interaction sites. This process is repeated for each individual probe or interaction type.

Following clustering of individual interaction types, clusters within the binding site of interest are selected for pharmacophore modeling. Due to the multiple potential interactions of isopropyl alcohol and n-methylacetamide, hydrogen-bond donor and acceptor clusters are manually selected for inclusion in the pharmacophore model by visualizing the surrounding protein structure and nearby high occupancy water molecules. The selected clusters are then consolidated, so that overlapping clusters are converted into a single pharmacophore feature. For example, aromatic and hydrophobic clusters occupying the same site would be assigned to a joint "aromatic|hydrophobic" interaction type. Overlapping clusters were defined as clusters having centers (local maxima) within 2.5 Å (for hydrophobic and aromatic clusters) or 1.5 Å (for hydrogen-bond donors or acceptors). The values of 2.5 Å and 1.5 Å are the approximate width of a pyrimidine probe and approximate Van der Waals radius of oxygen or nitrogen atoms, respectively. The larger value for aromatic and hydrophobic interactions accounts for the fact that these occupancies were determined based on center of mass rather than individual atom positions like hydroxyl groups. Local maxima separated by less than these distances would thus be considered to be occupying the same site. Pharmacophore features arising from only one probe type are represented using the highest occupied point as the center, with a radius given by the RMSD of all other points within the cluster to the center. Features that contain multiple probe types are represented as joint features using the average of all points as the center, with the radius given by the RMSD of all other points to this center. This process is shown in **Figure 5.3**.

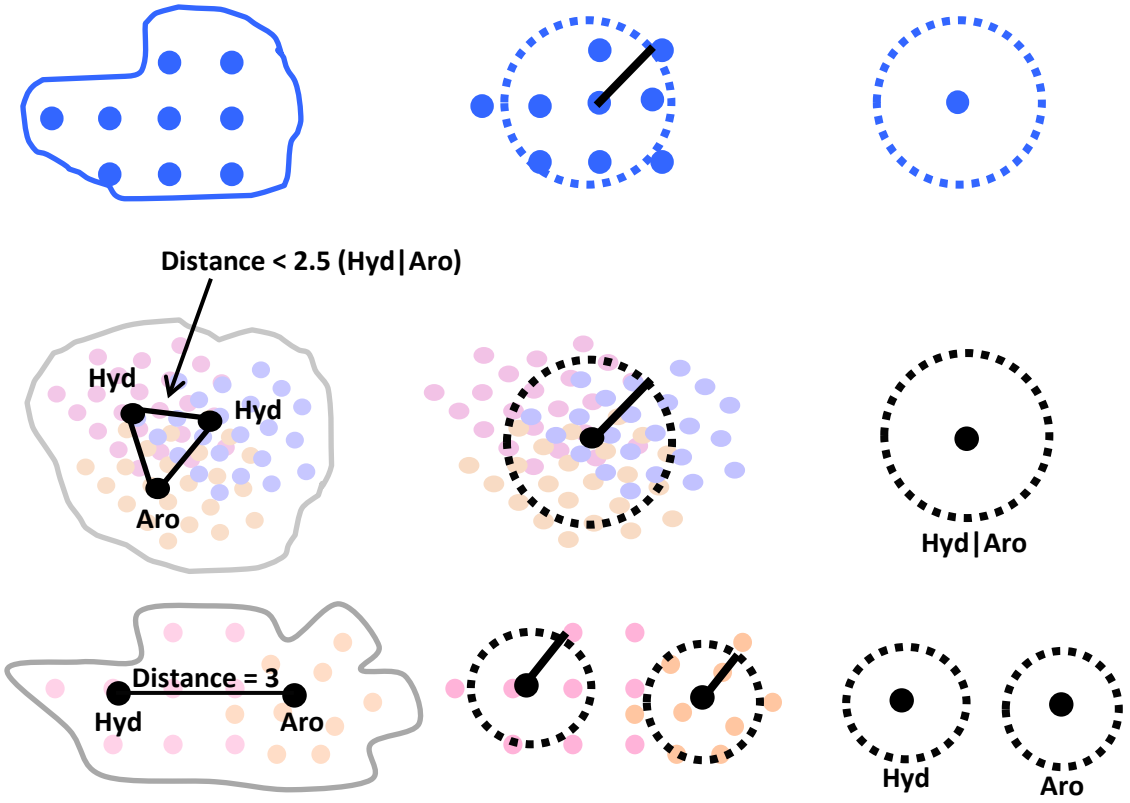


Figure 5.3: The DBSCAN algorithm is used to identify clusters of highly occupied grid points. Top) For each cluster of probe density, the highest occupied point is selected as the center and the RMSD of every other point to the center is calculated, to yield the radius of the pharmacophore feature. Middle) When multiple probes overlap within the specified cutoff, the average of all grid points within the cluster is used to define the center of the pharmacophore feature, and the radius is determined from the RMSD of all points to this center. Bottom) Maxima separated by a greater distance than the cutoff have minimal overlap, and are more appropriately represented as separate features.

Virtual Screening Procedure

Pharmacophore features are joined into a pharmacophore model and converted into the proper format for use with the program MOE and the PCH pharmacophore scheme[30]. Pharmacophore models of ABL Kinase were screened against all co-crystallized ligands of ABL (n=13)[230-240] and the DUD-E ABL1 Kinase final decoys set (n=10,750)[241]. Pharmacophore models of the putative allosteric sites in Src Kinase were screened against the ChemBridge CORE and EXPRESS libraries (Accessed July 2016: ChemBridge Corp., San Diego, CA), the Maybridge

Fragment and Hit finder libraries (Accessed July 2016: Maybridge, Thermo Fisher Scientific), Leicestershire, UK), and the ZINC Leads Now and Frags Now libraries[9, 242]. MOE allows for the stringency of pharmacophore models to be adjusted by using a multiplicative factor to increase the pharmacophore feature radii or to allow for partial matches. For ABL kinase, 1x to 2x radii were tested, for partial matches of 6-13 (of 13 total) possible pharmacophore features. Following previous work by our group[75, 243], partial matches were allowed to hit any of the possible features. Screening in this manner allows for the greatest number of potential compounds to be identified. As the occupancy cutoff used for clustering of Src was higher than that of ABL, pharmacophore features were smaller on average than those of ABL. To account for this, larger multiplicative factors were tested when screening the pharmacophore models. For Src kinase, 1, 1.33, 1.67, 2, and 2.33x radii were tested along with partial matches of 6-7 (of 7 total) possible pharmacophore features for the SH2-kinase interface site and 7-8 (of 8 total) possible pharmacophore features for the SH3-kinase interface. The coordinates and 1x radii of all pharmacophore models are given in the supplementary information.

5.4 Results and Discussion

Validation of Pharmacophore Protocol

In order to validate the MixMD pharmacophore creation protocol, pharmacophore models for the active site of ABL kinase were created. As shown in **Figure 5.4**, MixMD simulations correctly identify the active and allosteric sites of ABL kinase, as previously noted[51]. The grid points falling within this region were selected for conversion into a pharmacophore model, as shown in **Figure 5.5**. The observed occupancy in the ABL kinase active site spans a wide area and includes a number of potential interactions, resulting in 13 total pharmacophore features. However, it is unlikely that a single ligand could make all of these interactions at once. Indeed, virtual screening in MOE against known actives and the DUD-E ABL kinase decoy set requiring matches of 10-13 pharmacophore features yielded no hits. Requiring 6-9 pharmacophore features yielded a number of ligands that satisfied the requirements, shown in **Table 5.1**. As expected, increasing the pharmacophore radii by a multiplicative factor yielded a

greater number of matching ligands. A larger proportion of known active compounds satisfied the pharmacophore models than did inactive compounds, as shown in **Figure 5.6**. Models were further compared by calculating each model's deviation from a perfect model (ie. distance from the model data point to the ideal model having perfect sensitivity and specificity at 100% of actives, 0% of inactives)[243]. Based on this metric, pharmacophore models based on eight pharmacophore features had the best performance. Requiring eight pharmacophore features with 1.5x radii found 76.9% of active compounds and 19.1% of inactives. Using seven pharmacophore features with 1.1x radii performed similarly, finding 69.2% of known active compounds and only 11.4% of inactives. Increasing the radii to 1.3x and requiring a match to seven pharmacophore features identified 92.3% of active compounds, but had worse specificity, with 39.1% of inactive compounds satisfying the pharmacophore model. In order to perform the most unbiased test of the MixMD pharmacophore method, models were allowed to match any of the pharmacophore elements, as long as the required number was met. Including domain-specific knowledge to set a subset of required elements (such as the aromatic|hydrophobic regions on either side of the activation loop, or specific hydrogen-bonding interactions) would likely have improved the models' performance, but would bias the results. Regardless, the MixMD pharmacophore method successfully identified known active compounds from a large set of inactive compounds.

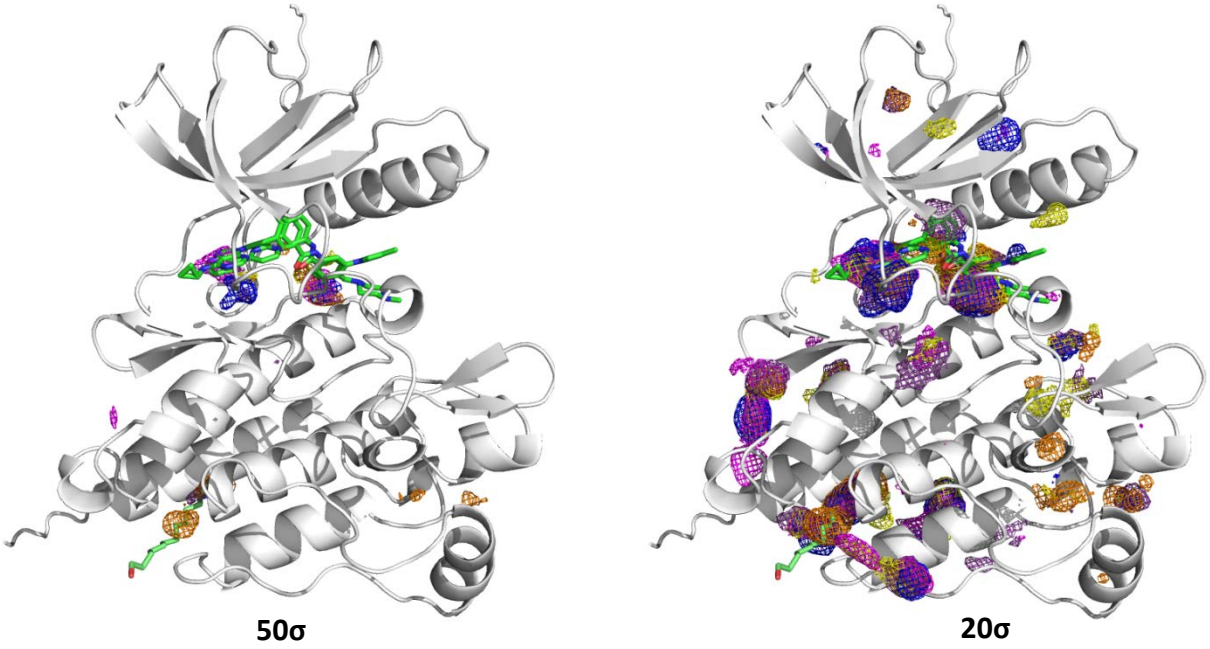
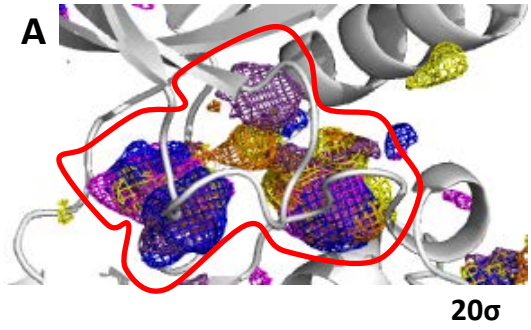
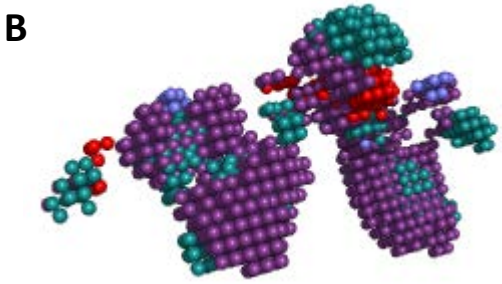


Figure 5.4: MixMD occupancy for acetonitrile (orange), imidazole (purple), isopropyl alcohol (blue), n-methylacetamide (yellow), and pyrimidine (magenta). The active and allosteric sites can be identified by the surrounding probe density, initially seen at very high occupancies (left). Visualizing the probe density at medium occupancy levels shows the extent of the binding site and full range of potential interactions (right). Ligands are shown in green for reference (PDB: 3KFA[235], 3MS9[236]), but were not included in the simulations.





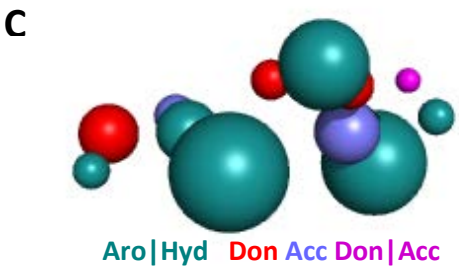


Figure 5.5: A) The region of ABL kinase mapped with the highest occupancy of MixMD density was selected for pharmacophore modeling. B) The occupancy for each interaction type was counted for each grid point. C) Grid points are clustered into pharmacophore features. Coordinates and radii are given in **Table 5.3**.

Pharmacophore Model Performance for ABL Kinase Active Site

Num. Features	9		8		7		6	
Radii	Actives	Inactives	Actives	Inactives	Actives	Inactives	Actives	Inactives
1x	0.0%	0.0%	0.0%	0.1%	53.8%	4.5%	92.3%	45.5%
1.1x	0.0%	0.0%	7.7%	0.5%	69.2%	11.4%	92.3%	64.5%
1.2x	0.0%	0.0%	7.7%	1.7%	69.2%	23.4%	100.0%	78.3%
1.3x	7.7%	0.1%	38.5%	4.6%	92.3%	39.1%	100.0%	86.7%
1.4x	15.4%	0.6%	53.8%	10.8%	-	-	-	-
1.5x	23.1%	1.9%	76.9%	19.1%	-	-	-	-
1.6x	30.8%	4.2%	-	-	-	-	-	-
1.7x	38.5%	8.1%	-	-	-	-	-	-
1.8x	46.2%	12.9%	-	-	-	-	-	-
1.9x	46.2%	19.1%	-	-	-	-	-	-
2x	61.5%	25.7%	-	-	-	-	-	-

Table 5.1: Percentage of tested compounds satisfying the pharmacophore model. Actives were taken from co-crystal structures of ABL in the protein databank (n=13) [230-240] and inactives were taken from the DUD-E ABL-1 kinase final decoy set (n=10,750) [241].

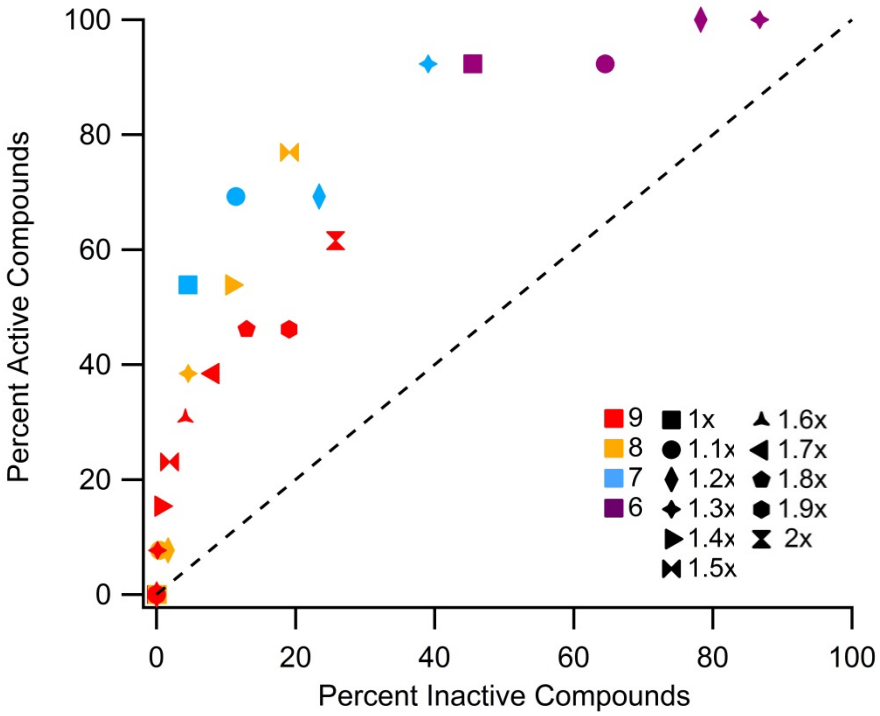


Figure 5.6: Percent of active compounds (n=13) satisfying the pharmacophore model of the ABL kinase active site relative to the percentage of inactive compounds (n=10,750). Pharmacophore models requiring 6-9 matches with 1-2x radii were tested.

Prospective Screening of Src Kinase

MixMD simulations of Src Kinase identified two pockets of density at the SH-2 and SH-3 kinase interfaces, as shown in **Figure 5.7**. Both sites were among the top ranked sites by maximal occupancy of probe molecules. Based on this ranking and their presence along the site of an important stabilizing interaction, we hypothesized that ligands targeting these sites may be able to stabilize the closed conformation of Src Kinase. To our knowledge, there are no known inhibitors targeting these sites. Pharmacophore models were created for the SH-2 and SH-3 sites separately, following the same clustering procedure used for ABL kinase. Screening the models against the ChemBridge and Maybridge libraries yielded a number of hits, as shown in **Table 5.2**. The hit rates for each pharmacophore model were compared in order to select a tractable number of compounds for purchase and experimental testing. As the stringency of the pharmacophore model is adjusted by incrementally increasing the radii, an increasing number of hits are obtained. Likewise, decreasing the number of required pharmacophore elements also increases the number of hits. As seen in the validation studies of ABL kinase, the appropriate pharmacophore models have a moderate level of stringency to successfully identify active compounds, without matching an overly large number of inactive compounds. Based on this, we selected the pharmacophore model requiring a match to all of the pharmacophore elements at 1.67x radii for the SH-2 kinase interface (87 hits) and 2x radii for the SH-3 kinase interface (45 hits) for subsequent testing. The selected pharmacophore models were then screened against the ZINC Leads Now and Frags Now subsets. This resulted in an additional 132 matching compounds for the SH2-kinase site and 35 for the SH3-kinase site. A subset of these compounds is currently being experimentally tested, using a protease accessibility assay, to determine each compound's effect on the global conformation of Src. This was developed by Matthew Soellner and coworkers at the University of Michigan, and is based on the exposure of the region linking the kinase and SH domains[244]. Compounds that do not induce the closed state allow for cleavage of this region by thermolysin. The ratio of open to closed states can then be measured over time from the band intensity of each product on an SDS gel. All experimental testing is being performed by our collaborators in the Soellner lab.

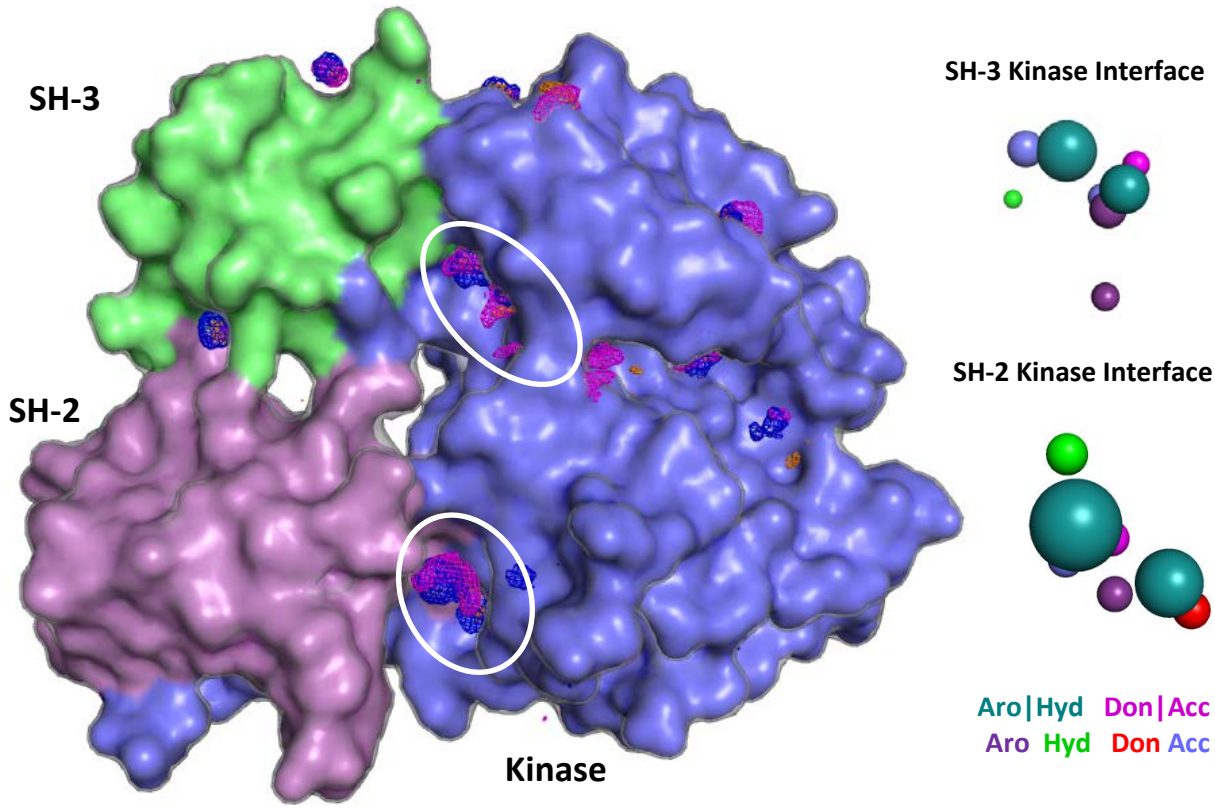


Figure 5.7: Left) MixMD density is shown for acetonitrile (orange), isopropyl alcohol (blue), and pyrimidine (magenta) contoured at 20 σ. The SH-2 and SH-3 domains form two pockets with the kinase interface, which ranked among the top sites (circled) by MixMD probe occupancy. Right) Pharmacophore models for the SH-2 and SH-3 kinase interfaces of Src. Spheres are colored according to the pharmacophore feature type. Coordinates and radii of the pharmacophore features are given in **Tables 5.4** and **5.5**.

Src Kinase Virtual Screening Results

		SH-2 Kinase Interface		SH-3 Kinase Interface	
Num. Features		7	6	8	7
Radii	1x	1	1035	0	0
	1.33x	7	-	0	29
	1.67x	**87**	-	2	1065
	2x	430	-	**45**	-
	2.33x	2354	-	466	-

Table 5.2: Matching compounds from screening the ChemBridge and Maybridge libraries against the pharmacophore models. Compounds were required to hit either all or all but one of the possible pharmacophore features. Bolded numbers indicate the pharmacophore models selected for further testing.

5.5 Conclusions

Using the procedures described herein, MixMD occupancy maps can be converted to pharmacophore models for virtual screening with MOE. Validation on ABL kinase showed a good ability of our models to correctly identify known active compounds from sets of inactive compounds. For ABL kinase, the extensive mapping by the probe molecules within the active site yielded a large number of pharmacophore features that were impossible to satisfy simultaneously. Screening using a reduced subset of the total features successfully identified known inhibitors, but also matched some of the inactive compounds. Increased specificity could potentially be achieved by manually selecting pharmacophore features for inclusion based on knowledge of known active compounds and their interactions with the binding site. Since the goal of this study was to validate the pharmacophore generation protocol rather than to screen for new inhibitors of ABL, this was not tested. Similarly, pharmacophore features could be selected based on the observed occupancy at each site during the simulation, so that high affinity sites are required features in the final pharmacophore model. This would likely help to bias the hit compounds towards those with higher potential affinities. Specificity may also be increased by incorporating additional shape information from the occupancy maps of the MixMD simulations. Pharmacophore models are typically represented using spheres centered at some point, but it is compelling to think that screening into the occupancy maps directly would incorporate more specific interaction preferences and lead to better distinguishing active and inactive compounds.

As screening of the compounds satisfying the pharmacophore models of the SH-2 and SH-3 kinase interfaces in Src are not complete, it is not known if any active compounds will be identified. Admittedly, this is a difficult task, as no ligands targeting these sites have been previously discovered and so it is unknown if ligands binding to this site are even capable of stabilizing the closed conformation. Nevertheless, we have developed a framework for conversion of MixMD results into pharmacophore models, which enables future prospective applications of the method.

5.6 Supplementary Information

ABL Kinase Active-Site Pharmacophore Model

Feature Type	x	y	z	r
Aromatic|Hydrophobic	5.3272	-5.0214	-1.1163	1.3145
Aromatic|Hydrophobic	-0.1101	-1.9205	9.8228	2.5990
Aromatic|Hydrophobic	7.8438	-2.8039	-0.9048	2.3757
Aromatic|Hydrophobic	1.8230	-3.1502	12.8002	0.8922
Aromatic|Hydrophobic	5.6607	-6.3754	5.6796	2.0000
Aromatic|Hydrophobic	7.2642	-4.2688	-6.0361	0.7080
Acceptor	1.6400	-3.6080	7.9810	1.5700
Donor	5.6980	-5.6720	-4.4910	1.1900
Donor	4.6580	-6.0810	5.5090	0.7700
Acceptor	3.8690	-5.9630	-0.8630	0.9200
Donor	4.3920	-6.4490	3.6360	0.8900
Donor	4.1758	-5.3850	7.4456	1.0892
Donor|Acceptor	5.8791	-4.7948	9.3955	0.5605

Table 5.3: Feature type, coordinates, and radius for all potential features in the pharmacophore model of the ABL kinase active-site.

Src Kinase SH2-Kinase Interface Pharmacophore Model

Feature Type	x	y	z	r
Anion	5.5230	-5.6350	16.7550	0.9500
Anion	6.0370	-5.6670	19.2520	0.8700
Cation	9.0930	-2.7200	17.7590	1.0000
Cation	8.6420	-0.7180	18.7790	0.6100
Aromatic	8.2500	-6.7500	15.7500	0.6700
Hydrophobic	10.7840	-6.9740	10.4210	0.7100
Aromatic|Hydrophobic	8.0313	-4.2384	13.5361	1.8173
Aromatic|Hydrophobic	3.0393	-4.6260	16.5623	1.4948
Acceptor	8.3230	-3.9700	14.9010	0.7100
Donor	1.8200	-4.8830	17.9230	0.8500
Donor|Acceptor	5.9943	-4.1058	14.3220	0.7301

Table 5.4: Feature type, coordinates, and radius for all potential features in the pharmacophore model of the SH2-Kinase interface in Src Kinase.

Src Kinase SH3-Kinase Interface Pharmacophore Model

Feature Type	x	y	z	r
Anion	5.8340	-6.9120	-10.7590	1.2300
Anion	7.8500	-6.4640	-9.7670	0.5000
Anion	9.3950	-6.5650	-1.7760	1.9000
Cation	5.9760	-1.4480	-6.7140	0.7100
Aromatic	5.7500	-5.2500	-10.2500	0.8800
Aromatic	8.7500	-8.2500	-7.2500	0.5900
Hydrophobic	6.0140	1.3820	-9.7700	0.5000
Aromatic|Hydrophobic	5.4768	-1.8869	-12.6331	1.5279
Aromatic|Hydrophobic	7.6836	-7.8453	-11.6304	1.0508
Acceptor	6.7650	-0.2800	-12.5710	0.9200
Acceptor	4.7170	-3.7590	-10.5830	0.6500
Donor|Acceptor	4.0605	-5.2698	-12.3397	0.6608

Table 5.5: Feature type, coordinates, and radius for all potential features in the pharmacophore model of the SH3-Kinase interface in Src Kinase.

Chapter 6. MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations

6.1 Abstract

Mixed-solvent molecular dynamics (MixMD) is a cosolvent simulation technique for identifying binding hotspots and specific favorable interactions on a protein's surface. MixMD studies have the ability to identify these biologically relevant sites by examining the occupancy of the cosolvent over the course of the simulation. However, previous MixMD analysis required a great deal of manual inspection to identify relevant sites. To address this limitation, we have developed MixMD Probeview as a plugin for the freely available, open-source version of the molecular visualization program PyMOL. MixMD Probeview incorporates two analysis procedures: 1) to identify and rank whole binding sites and 2) to identify and rank local maxima for each probe type. These functionalities were validated using four common benchmark proteins, including two with both active and allosteric sites. In addition, three different cosolvent procedures were compared to examine the impact of including more than one cosolvent in the simulations. For all systems tested, MixMD Probeview successfully identified known active and allosteric sites based on the total occupancy of neutral probe molecules. As an easy-to-use PyMOL plugin, we expect that MixMD Probeview will facilitate identification and analysis of binding sites from cosolvent simulations performed on a wide range of systems.

6.2 Introduction

First introduced in 2009[72], hotspot mapping with molecular dynamics (MD) simulations of small molecule probes and water is being increasingly applied towards the development of small molecule inhibitors. These cosolvent simulations provide two types of information. First,

when many probes map a location, it identifies binding sites on the protein's surface, including ligand binding sites, protein-protein interaction sites, and other biologically relevant interactions. Secondly, the functional groups on the individual probes identify functional sites on the protein's surface that favor specific interactions, which can be used to inform structure-based drug design efforts. Several cosolvent simulation methods have been introduced, as recently reviewed by Ghanakota and Carlson[62]. While these methods all utilize mixtures of small molecule probes and water, they have a number of differences regarding the specific probes used, the protocol for simulation, and the method of identifying and ranking the results. For example, some cosolvent methods have focused on the use of a single probe molecule per simulation while others have multiple probes run simultaneously. The MixMD method developed by our group previously utilized a layered setup of a single probe type and water in a 5%/95% v/v probe to water ratio[79]. Introducing charged probes required a transition to ternary solvent mixtures to balance the number of positive and negative charges within the system[51]. Other simulation methods, including the SILCS method[67] from the MacKerell group and cosolvent simulations by Bakan et al.[245], have utilized 4-7 different types of probe molecules within the same simulation. Simulations containing multiple probe types clearly require fewer simulations than comparable methods that simulate each probe separately, but the extent to which this influences the predicted binding sites is unclear.

Traditionally, hotspots have been identified by overlapping density from multiple probe molecules[43, 49, 51]. In our MixMD method, the occupancy of probe molecules is determined by overlaying the protein and solvent system with a grid and counting the number of times a probe molecule occupies each region. The occupancy is then converted into "σ units", expressed as the number of standard deviations away from the mean occupancy. This allows for the maps to be viewed at different occupancy contour levels, in an analogous way to crystallographic electron density. The resulting maps are visualized in PyMOL to identify the highest occupied sites comprised of multiple probe types. These regions, or hotspots, are then ranked by maximal occupancy[51]. When applied to seven test systems, this method successfully identified known biologically relevant sites on the basis of maximal occupancy[51]. However,

manually inspecting every probe map at multiple occupancy contour levels for every system is tedious and time-consuming, thereby limiting the number of systems that can be studied.

Other approaches have identified binding sites by converting the probe occupancies into theoretical binding affinities. In the SILCS method[67] and the method by Bakan et al.[245], the binding affinity at a specific grid point is calculated from the Boltzmann relationship:

$$\Delta G_i = -RT \ln(\frac{O_i}{O_{bulk}}) \quad (1)$$

where O_i is the occupancy at grid point *i*, O_{bulk} is the expected occupancy in bulk solvent, and T is the temperature. In the SILCS method, these energies are referred to as grid free energies, and they can be used to visualize predicted affinities on the surface of the protein or may be used to determine the theoretical binding affinity of a ligand having atoms at point i[67]. In the approach used by Bakan et al., distinct interaction sites are identified, and the lowest energy point, calculated from Equation 1, is selected to represent the site[245]. Nearby sites are merged and the energies are summed to yield theoretical affinities for each region. The affinities are then used to rank the "druggability" of each site. This approach was used successfully to identify known binding sites for five systems, and to rank potential binding sites within each system by the maximum predicted affinity[245].

While Equation 1 is straightforward to use, there are some inherent limitations in the calculation of binding affinities at the level of sub-atomic grid points using data from simulations of whole probe molecules. The binding affinity of a probe molecule is dependent on the contributions of every atom within the probe. For example, in the case of isopropyl alcohol, the hydroxyl group may be making hydrogen bonding interactions, while the methyl groups are making hydrophobic interactions. Partitioning the binding affinities calculated from the entire probe molecule's occupancy down to the grid point level neglects to consider these effects. Instead, we have focused on the analysis of overall occupancy of the probe molecules as a whole. Using a clustering method to identify separate regions on the protein's surface, we

calculate the total occupancy of probe molecules for each site across all simulations. This identifies the regions that are highly occupied by multiple probe types across multiple simulations.

To facilitate application of our MixMD method, we have developed a plugin, which we call MixMD Probeview, for use with the freely available open-source version of PyMOL[100]. Requiring only a PDB-formatted file containing grid points and associated occupancies from a set of cosolvent simulations (easily obtained by post-processing of trajectories with AmberTools[187]), MixMD Probeview identifies binding sites composed of multiple probes as well as local maxima for individual probes. We have validated this method on four systems (including two with allosteric sites), using data taken from more than 2 μs of simulation time per system. Simulations were performed for multiple solvent setup procedures, including both solvents alone (ie. a single probe and water) and in several combinations (ie. 2 or more probe types and water). This allowed us to verify the ability of MixMD Probeview to identify binding sites for a range of systems and cosolvent procedures. Additionally, since simulations were completed for both individual probes and probes run in several different mixtures, we were able to compare the resulting probe occupancy and binding site prediction for different simulation methods. For each system and solvent mixture, the simulations were analyzed at two levels. The first being the ability to correctly predict and identify biologically relevant regions as highly ranked hotspots, and the second being the agreement in functional group mapping between individual and combined probe simulations.

6.3 Methods

ABL kinase (PDB:3KFA)[235], Androgen receptor (AR, PDB:2AM9)[246], β-secretase (BACE, PDB:1W50)[174], and dihydrofolate reductase (DHFR, PDB:1DG8)[176] were selected as test systems. These proteins are commonly used benchmark systems and include systems with allosteric sites to provide a thorough test of MixMD Probeview's ability to predict binding sites. All ligands and water molecules in the crystal structures were removed, with the exception of the NADPH

cofactor in DHFR which was retained and modeled using the parameters developed by Ryde[183, 184]. Hydrogens were added and asparagine, glutamine, and histidine positions were optimized using MolProbity and the Protonate 3D tool in MOE[30, 182]. For each system, probes were run individually ("solo") or in one of two combined sets, given in **Table 6.1**. Portions of these simulations were completed previously by our group[51]. Solvent mixtures were chosen to minimize the need for two probes to compete for mapping the same type of interaction with the protein surface. For example, pyrimidine and imidazole are both aromatic probes and would be expected to occupy many of the same sites. For this reason, none of the probe mixtures include both pyrimidine and imidazole. In each case, a 5%/95% v/v ratio of probe molecules to TIP3P[155] water was maintained, with the 5% of probe molecules split evenly between probe types.

<table>
<tr><th>Solo</th><th>Combination A</th><th>Combination B</th></tr>
<tr><td>Acetonitrile (ACN)</td><td rowspan="2">Acetonitrile
+ Isopropyl Alcohol</td><td rowspan="2">Acetonitrile
+ Isopropyl Alcohol
+ Imidazole</td></tr>
<tr><td>Isopropyl Alcohol (IPA)</td></tr>
<tr><td>Imidazole (IMI)</td><td rowspan="2">Imidazole
+ N-methylacetamide</td></tr>
<tr><td>N-methylacetamide (NMA)</td><td rowspan="3">N-methylacetamide
+ Pyrimidine
+ Methylammonium
+ Acetate</td></tr>
<tr><td>Pyrimidine (PYR)</td><td rowspan="2">Pyrimidine
+ Methylammonium
+ Acetate</td></tr>
<tr><td>Methylammonium (MAI)
+ Acetate (ACT)</td></tr>
</table>

Table 6.1: The boxes indicate the probe mixtures used for each set of simulations. The solo probes were all run as a single probe in combination with water, except for methyalmmonium and acetate, which must be run together to achieve an overall neutral charge.

The simulations were initiated using a layered setup, with probe molecules placed around the protein, followed by a box of water to achieve the desired concentration. This setup was chosen to facilitate probe sampling at lower concentrations, consistent with previous development of the MixMD method[79]. The tleap module of AmberTools12 or 14[185, 187] was used for system setup, with the FF99SB[247] force field and previously developed solvent parameters[51, 73]. The systems were initially minimized, followed by heating to 300 K with restraints on the protein. The restraints were then gradually removed as the systems were equilibrated. For each system and solvent type, 10 simulations of 20 ns production time with a 2 fs timestep were completed with AMBER12 or 14[149, 185, 187, 226, 227]. Proper solvent behavior

over the course of the simulation was verified using radial distribution functions calculated using the cpptraj module in AmberTools14[187]. Following simulation, the last 10 ns of each trajectory were aligned, and the occupancy of the center of mass of each probe molecule was calculated on a 0.5 x 0.5 x 0.5 Å grid using an in-house modified version of the cpptraj module in AmberTools14[157, 187]. The modification to cpptraj was necessary to allow for center-of-mass based occupancies to be calculated.

Our PyMOL plugin, MixMD Probeview, was used for the analysis of the occupancy grids. The plugin consists of two analysis procedures: 1) to identify and rank whole binding sites and 2) to identify and rank maxima of each probe type. MixMD Probeview is written in Python and uses the scikit-learn package for clustering[248]. In order to identify whole binding sites on the protein's surface, the DBSCAN clustering algorithm was used. This algorithm is capable of identifying density connected regions of any shape or size and does not require a predefined cluster size or number of clusters[229]. DBSCAN clustering relies on three parameters: 1) a cutoff to determine which grid points to cluster, 2) epsilon (ε), the maximum distance by which two points can be separated and still be considered within the same cluster, and 3) the minimum number of points that must be contained in a cluster to consider it a valid cluster. Clusters are created by grouping all points that are reachable within the epsilon distance, and containing at least the minimum number of points. An example of the DBSCAN clustering process is shown in **Figure 6.1**. In practice, this allows for the automated identification of clusters of probe occupancy from either overlapping or adjacent grid points. The DBSCAN algorithm is particularly useful for identifying ligand binding sites because of its requirement for connected regions of density, thereby identifying sites that could be connected within the span of a few bond lengths. In the present study, grid points having greater than 10% of the maximum occupancy were used for clustering with a distance parameter of 3 Å. This is approximately the width of pyrimidine or twice the length of a carbon-carbon bond in ethane, and so would identify regions that could be connected within 1-2 bond lengths. The minimum number of points was set to 10 to remove small clusters from further analysis. Following clustering of the occupancy grid points, the resulting clusters can be ranked by either the maximum occupancy

found in the cluster or the total sum of occupancy within the cluster.

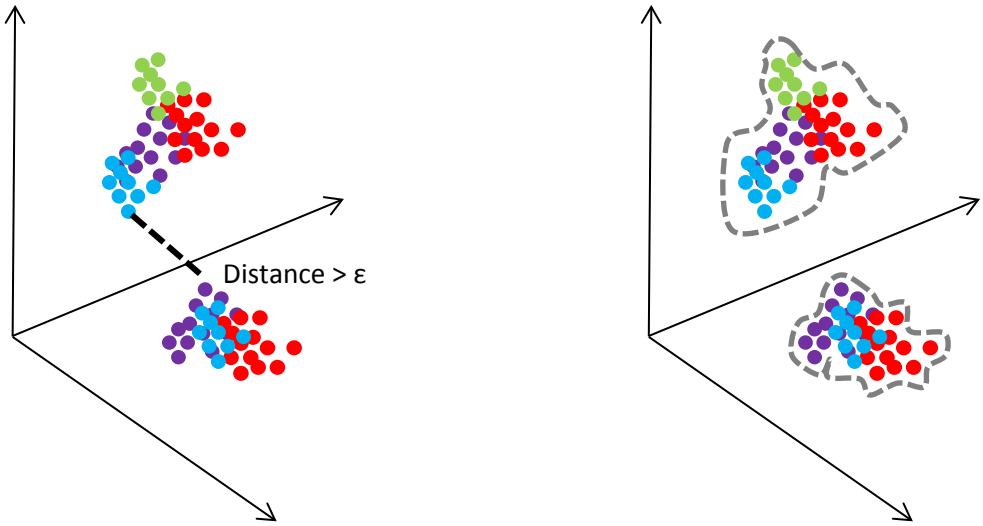


Figure 6.1: The DBSCAN clustering procedure identifies connected regions of probe density arising from multiple probe types (represented by different colors). Grid points within the distance parameter ε are grouped into the same cluster. The resulting clusters can then be ranked based on the probe occupancy within the cluster.

While the DBSCAN clustering algorithm is suitable for identifying binding sites, it is not capable of differentiating groups of points whose edges are adjoining, as frequently happens in regions adjacent to local maxima. In order to identify and rank favorable probe binding sites for individual probe molecules, the Mean Shift clustering algorithm was used[249]. The Mean Shift algorithm was chosen as it is capable of identifying arbitrary shapes and sizes of clusters from data points with varying density in 3-D space, making it ideally suited to finding clusters corresponding to local maxima from cosolvent simulations. In the Mean Shift clustering procedure, the distribution of data is represented by a kernel density estimate with bandwidth parameter h. An iterative process is then applied to the data to identify a local density gradient followed by a shift of the center of the kernel until the gradient of the density is zero, and the peak in the data is identified[249]. As depicted in Figure **6.2**, this clustering process identifies the highest occupied region as the center, with lesser occupied regions surround this point grouped into the cluster based on the observed spatial distribution. Larger bandwidth values will generate fewer, larger clusters while a smaller bandwidth value will give a greater number of small clusters. The clusters can then be ranked by the maximum occupancy within the cluster.

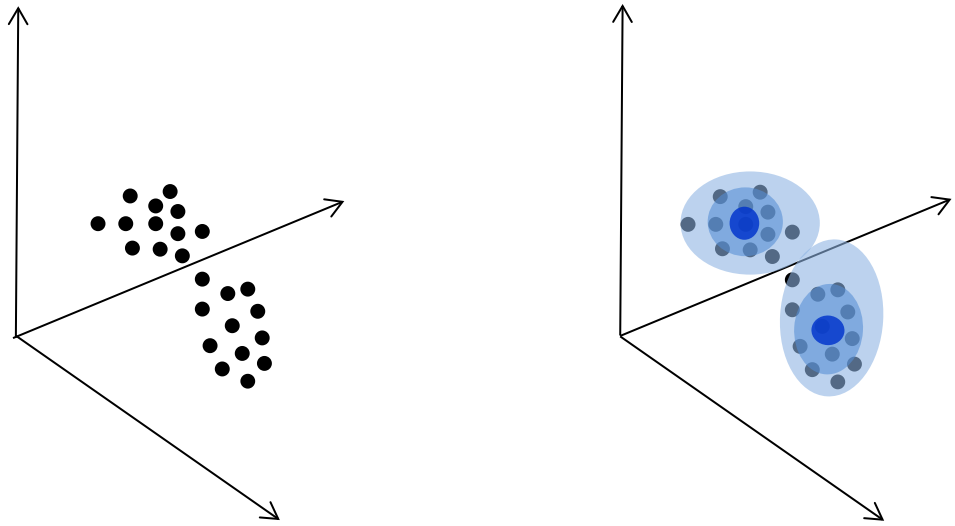


Figure 6.2: The mean shift clustering algorithm groups points based on their distribution in space. Densely occupied regions correspond to the center of a cluster (dark blue), while sparsely occupied regions indicate cluster edges (light blue).

6.4 Results and Discussion

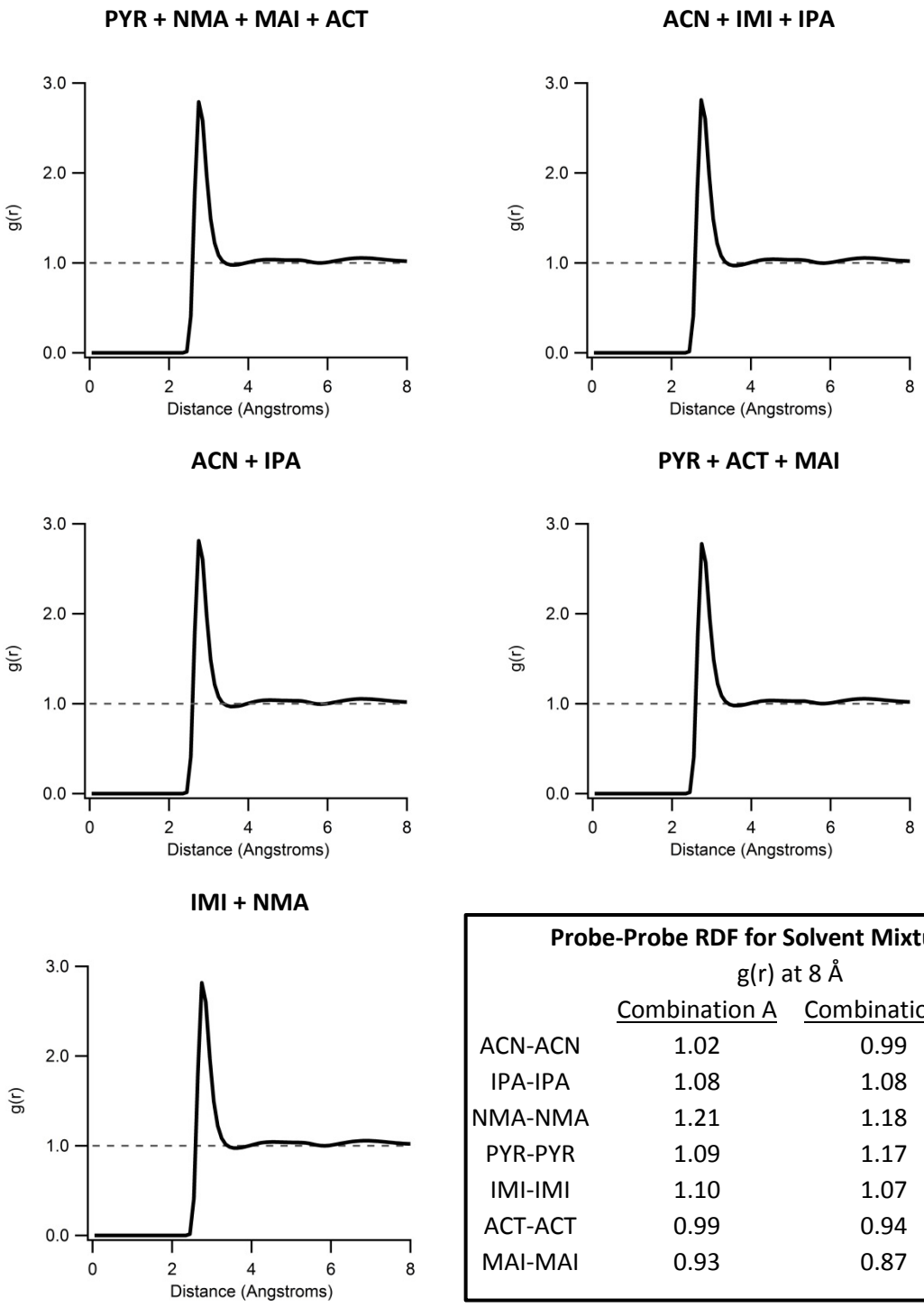
Validating Parameters for Combined Solvent Mixtures

Previous work by our group has highlighted the importance of validating solvent parameters prior to their use in cosolvent simulations[73]. Improper sampling of solvent positions would invalidate the underlying assumptions used when considering the ranking of observed probe occupancies. MixMD solvent parameters were previously shown to yield proper solvent behavior for runs of a single probe mixed with water, but could potentially have different behavior when used with combinations of solvents. As shown in **Figure 6.3**, all solvent mixtures showed proper mixing, with all g(r) values at 8 Å close to 1. Example values shown were taken from the production simulations of DHFR. This indicates that the solvent molecules were evenly mixed with no condensation into a separate phase from the water.

DBSCAN Clustering to Identify Binding Sites

Previous studies by our group established that biologically relevant sites could be identified based on maximum solvent occupancy in cosolvent simulations[51]. These simulations correctly identified the active and allosteric sites as being among the top ranked sites by occupancy. However, when solvent mixtures are used rather than single cosolvents some sites

that would normally be ranked as having maximal occupancy may have intermediate occupancy values because of multiple, exchanging solvent molecules. Ranking by maximal occupancy in these cases would favor sites that bind a single probe type tightly rather than those sites which bind multiple probe types tightly. To account for this, we have moved to ranking based on total occupancy within a region. Occupancies were generated based on the center of mass of each probe molecule so that each probe would contribute equally when the total (summed) occupancies were calculated. The rankings shown in the following sections were generated using our PyMOL plugin with the occupancies of all neutral probe molecules. Previous MixMD studies have shown the ability to identify most active sites using the occupancies of neutral probe molecules[51]. Charged probe molecules were included in each set of simulations, and yield additional insight into the binding preferences of each site.



Probe-Probe RDF for Solvent Mixtures

g(r) at 8 Å

	Combination A	Combination B
ACN-ACN	1.02	0.99
IPA-IPA	1.08	1.08
NMA-NMA	1.21	1.18
PYR-PYR	1.09	1.17
IMI-IMI	1.10	1.07
ACT-ACT	0.99	0.94
MAI-MAI	0.93	0.87

Figure 6.3: Radial distribution functions of the oxygen in water show expected behavior in all cases. Probe-probe radial distribution functions deviate slightly from 1, but are within the acceptable ranges previously established by our group. All values shown were taken from the production portion of the DHFR simulations.

ABL Kinase

Both active and allosteric ligands bind to ABL kinase, with varying specificity depending on ABL's conformational state. As shown in **Figure 6.4**, MixMD simulations identify both the active and allosteric sites as the top ranked sites for every solvent combination tested, though the ordering differed depending on the solvent set. Ligands that bind to the active site of the inactive, DFG-out form of ABL kinase (used to initiate the MixMD simulations) form interactions at the ATP binding site as well as the site that is occupied by phenylalanine in the DFG-in conformation[237, 250]. These two sites are encompassed by the MixMD identified binding site, which shows two areas of density connecting over the activation loop. As shown in **Figure 6.4**, there is a patch of highly occupied probe density at both the left and right sides of the active-site ligand, corresponding to the ATP and phenylalanine positions, respectively. Summing over the clustered grid points identifies the active site as having the highest total occupancy for both the individual probe and solvent combination B simulations. In the case of the simulations of solvent combination A, the left and right portions of the active site are broken up into two clusters, as they are separated by slightly more than 3 Å. This results in the active site being ranked second, behind the allosteric site. Including the second cluster at the left side of the active site would have ranked the active site as the highest occupied cluster. Regardless, the top two sites clearly have a greater degree of occupancy than other sites, as seen in the boxplot in **Figure 6.4**.

The second site identified by the MixMD occupancy corresponds to the allosteric site of ABL kinase. Allosteric ligands bind in the myristate pocket, near the C-terminus. In the autoinhibited form of ABL kinase, the C-terminus adopts a bent conformation, allowing the SH-2 and SH-3 domains to close against the adjacent kinase face[250]. Ligands binding to this site can act to stabilize the autoinhibited form of ABL (eg. GNF-2, PDB:3K5V)[251], or may block bending of the helix to stabilize the active conformation (eg. DPH, PDB:3PYY)[252]. Both allosteric activators and inhibitors occupy the myristate binding site, shown in dark blue in **Figure 6.4**. Activators form additional interactions to the left of this site, which effectively blocks helix bending. These additional interactions are replicated in the MixMD simulations, and correspond to the small, light blue cluster to the left of the dark blue allosteric site.

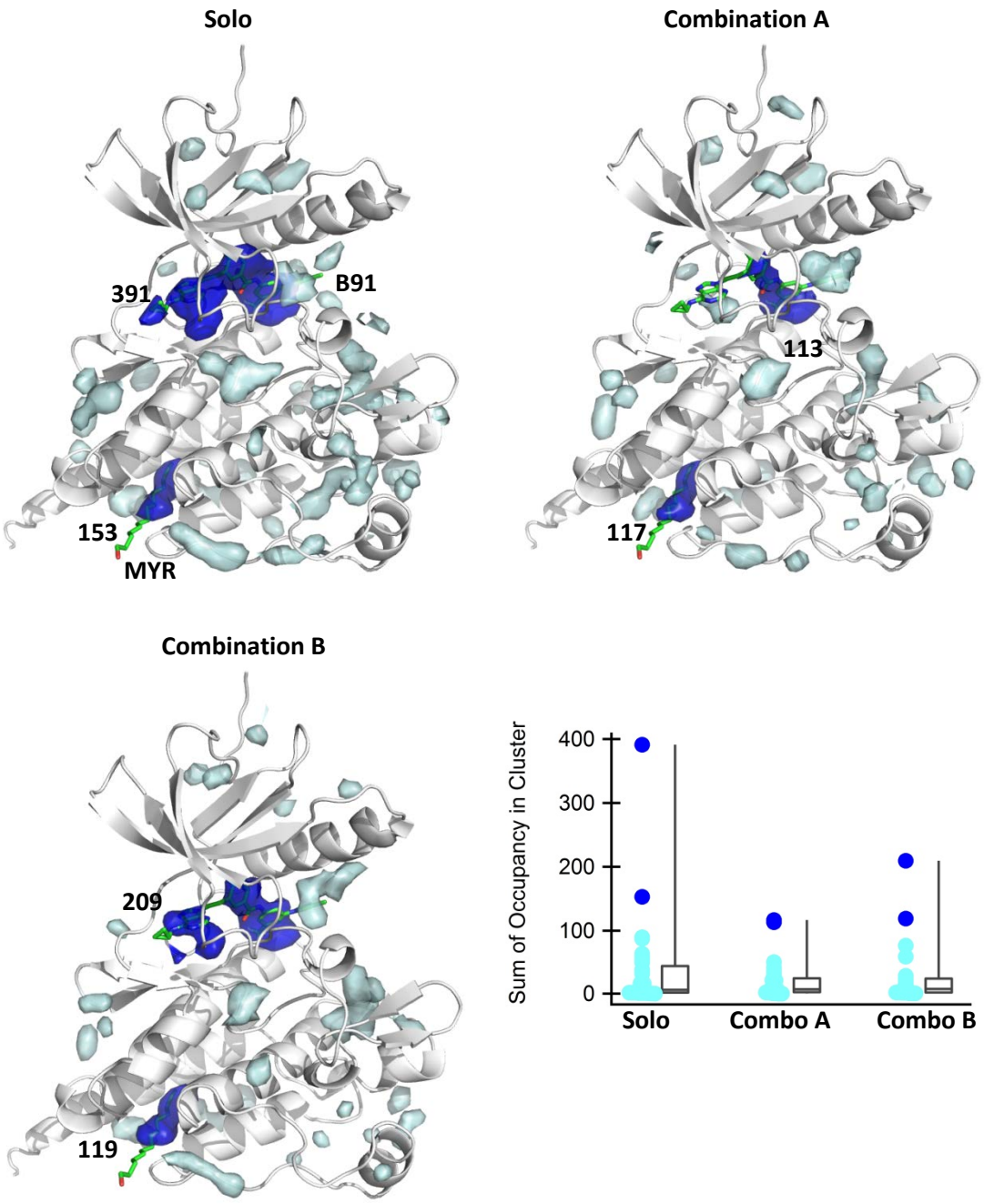


Figure 6.4: Cluster ranking by total occupancy for ABL kinase. The active site ligand B91 (PDB:3KFA)[235] and allosteric ligand (myristate, PDB:1OPJ)[250] are shown for reference. The top two sites for each solvent set are shown as dark blue clusters, with the total occupancy within these clusters given in bold. In every case, ranking by total occupancy identifies the active and allosteric sites as the highest ranked sites. The boxplot shows the distribution of total occupancies for each cluster and solvent set. The top two sites (corresponding to the active and allosteric sites) are noticeably higher in occupancy than the remaining clusters (light blue).

Androgen Receptor (AR)

Androgen receptor (AR) is a soluble steroid-type protein that acts as an intracellular transcription factor[253]. AR is stimulated by androgens (e.g., testosterone and 5α-dihydrotestosterone) which bind to the active site and regulate gene expression for male sexual characteristics. Both agonists and antagonists of AR have been developed to treat conditions such as hypogonadism and prostate cancer[253]. As shown in **Figure 6.5**, ranking by total occupancy from MixMD simulations successfully identifies the active site as the top ranked site in all three solvent sets.

AR also contains two allosteric sites, as shown in **Figure 6.5**. Ligands binding to these sites alter the receptor's conformation, and subsequently, its ability to bind to steroid receptor coactivator 2-3(SRC2-3)[254]. The inability to bind to SRC2-3 hinders the receptor's functionality, which ultimately diminishes the androgen response. These allosteric sites were identified in all three sets of simulations, but the ranking differed depending on the solvent set used. In the solo and solvent combination A simulations, the active site and two allosteric sites were the top three ranked sites. In the simulations of solvent combination B, the active site was ranked as number 1, but the two allosteric sites were ranked lower than one site that is a crystal packing interface. Comparing the distribution of occupancies among clusters, this discrepancy might be due to the smaller number of individual simulations for solvent combination B relative to the other solvent combinations. For each solvent set, 10 independent simulations of 20 ns are performed per solvent mixture. This results in 50 simulations (not including charged probes) for the solo simulations, 30 for solvent combination A, and 20 for solvent combination B. Averaging over a larger number of simulations might better distinguish functional binding sites from other easily desolvated sites on the protein's surface, as shown in the boxplot in **Figure 6.5**.

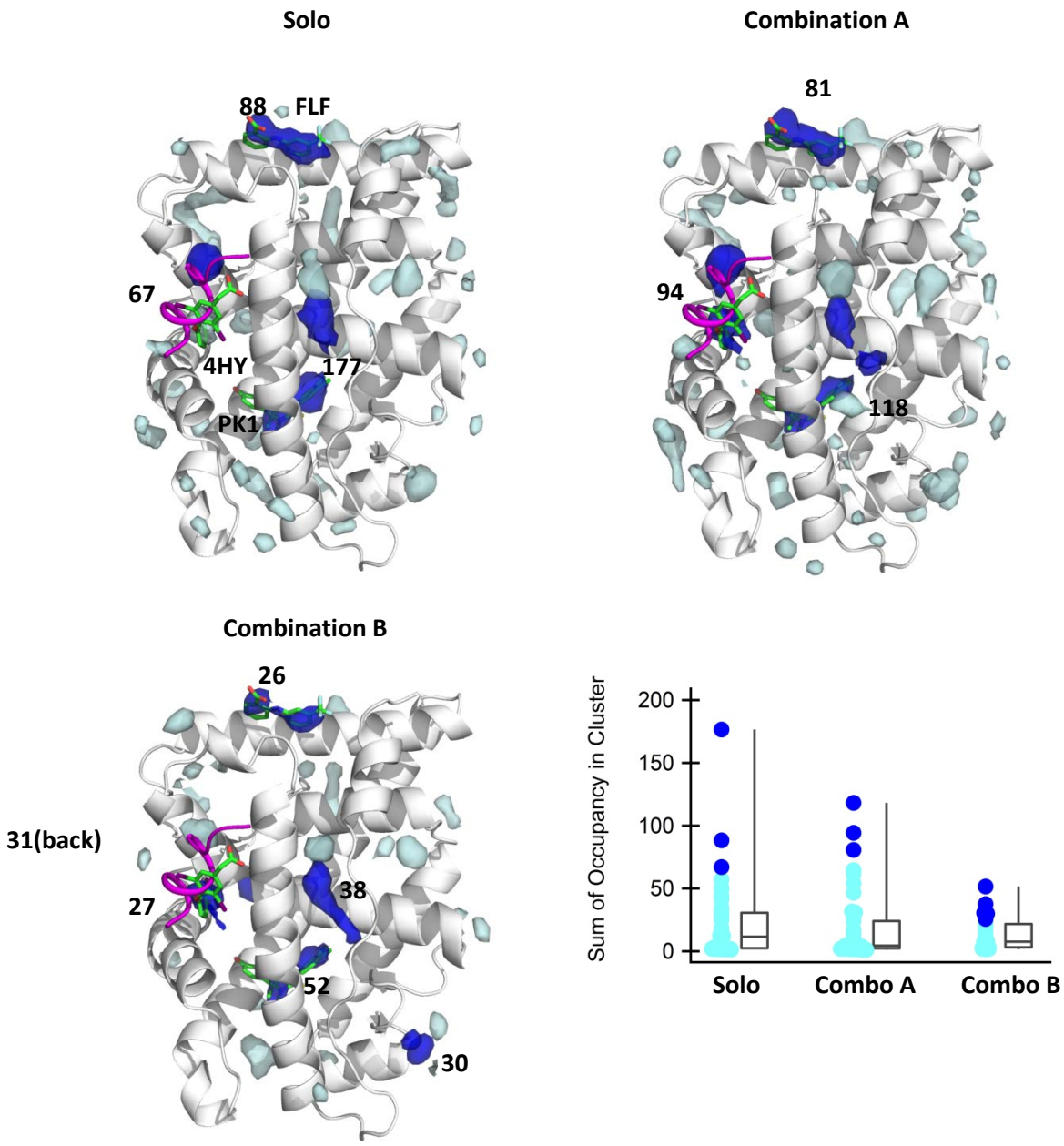


Figure 6.5: Cluster ranking by total occupancy for androgen receptor. The top ranked sites by occupancy are shown in dark blue, with the total occupancies for these clusters in bold. All other clusters are shown in light blue. Active (PDB:3V4A, PK1)[255] and allosteric (PDB:2PIU,4HY and PDB:2PIX, FLF)[254] ligands are shown for reference. The SRC-2 coactivator peptide is shown in magenta (PDB:2QPY)[254]. The active site is the top ranked site in all cases. In the solo and solvent combination A simulations, the two allosteric sites are the next highest ranked sites. However, in solvent combination B the total occupancies for the remaining sites are close together, making it difficult to discern the allosteric sites from ranking alone.

β-Secretase

BACE is responsible for cleavage of β-amyloid precursor protein[256]. The active site of BACE is a large cleft, containing a number of known subsites involved in ligand recognition[257-259]. Ligands do not have to make all of these interactions however, and effective ligands have been developed that bind within only a small region of the overall active site. For example, LY2811376 binds BACE with nanomolar affinity by engaging the catalytic aspartates and S1 and S3 subsites, and leads to decreased levels of Aβ in animals and humans[260]. MixMD simulations correctly identify these subsites, showing the highest levels of probe occupancy within the region occupied by LY2811376. As shown in **Figure 6.6**, MixMD identifies the active site cleft as the highest ranked site for every solvent set tested, though the spread of the clusters differs. The cluster from the solo simulations spans the largest area, with probe occupancy extending across the binding cleft. In the solvent combination A and B simulations, a smaller region is mapped, but this smaller region corresponds to the portion of the active-site known to be targetable by small, high-affinity inhibitors.

Dihydrofolate Reductase

Dihydrofolate reductase (DHFR) is an enzyme that catalyzes the transformation of dihydrofolate to tetrahydrofolate, which is utilized for purine and thymidylate synthesis. Since DHFR is the sole source of tetrahydrofolate, DHFR is a common therapeutic target for many antibiotics, autoimmune disorders, and cancers[261]. As shown in **Figure 6.7**, MixMD correctly identifies the active site as the top-ranked site for every solvent mixture. All ligands binding within the active-site of DHFR occupy a T-shaped cleft, which is identified as the most-highly occupied site in our simulations. Some ligands extend beyond this core area to make additional interactions. For example, methotrexate contains two carboxylate groups that bind at the very edge of the active-site region. Identification of binding sites was based on neutral probe occupancy, so these sites are not visible in **Figure 6.7**, but are seen as local maxima of acetate in **Figure 6.11**. This demonstrates the ability of MixMD to correctly identify the core active-site region as well as accessory sites that may be utilized by some ligands.

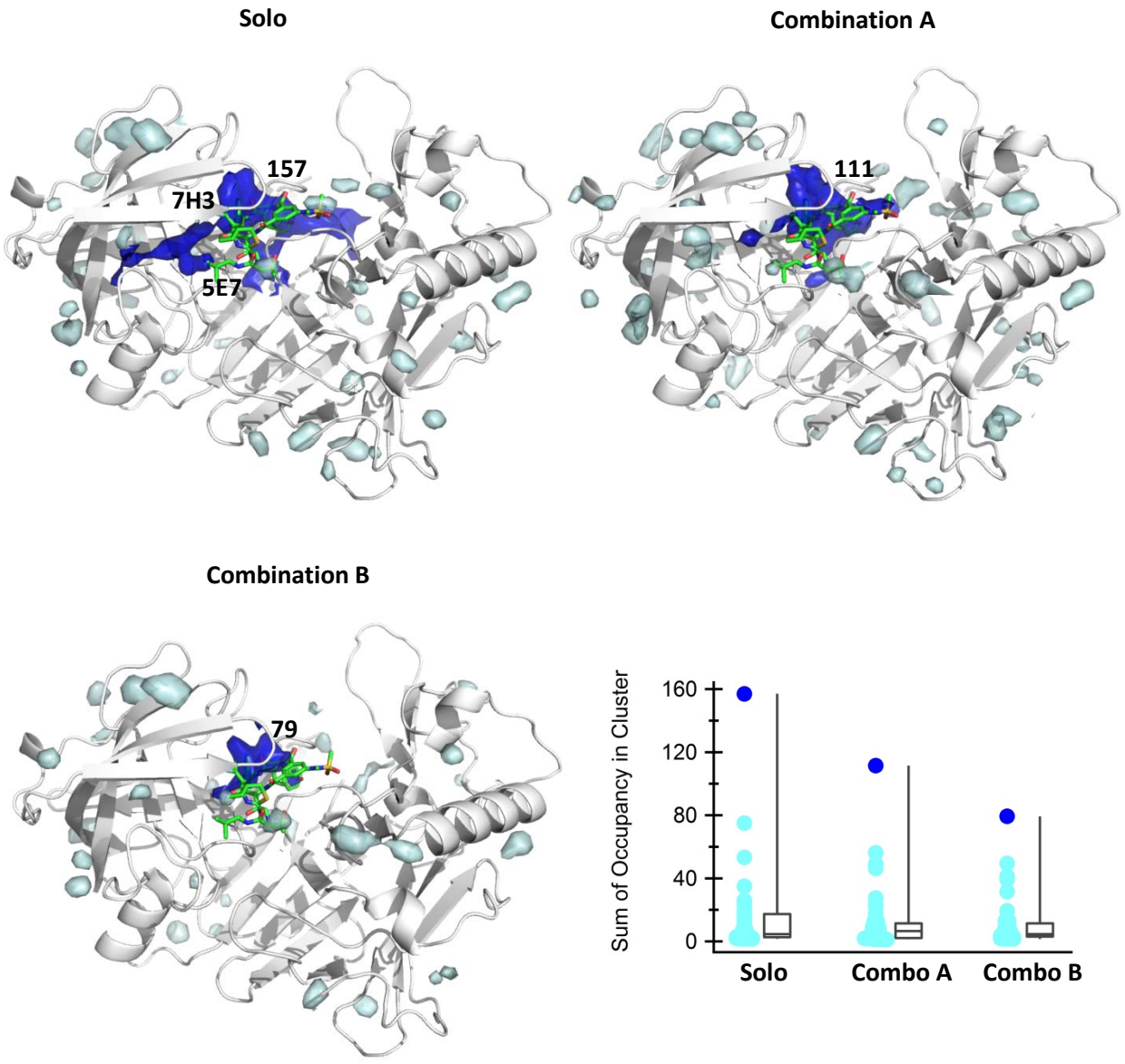


Figure 6.6: BACE contains an extended binding cleft, with inhibitors 7H3 (PDB: 5TOL)[262] and 5E7 (PDB:5DQC)[263] shown for reference. In every case, MixMD correctly identifies the active site as the region with the highest total occupancy, shown in dark blue. The total occupancies of the top clusters are given in bold, with the remaining clusters shown in light blue. The top cluster identified from solvent combinations A and B is smaller than that of the solo simulations, but overlaps with the subsites of BACE that have been targeted by small, high-affinity ligands[260].

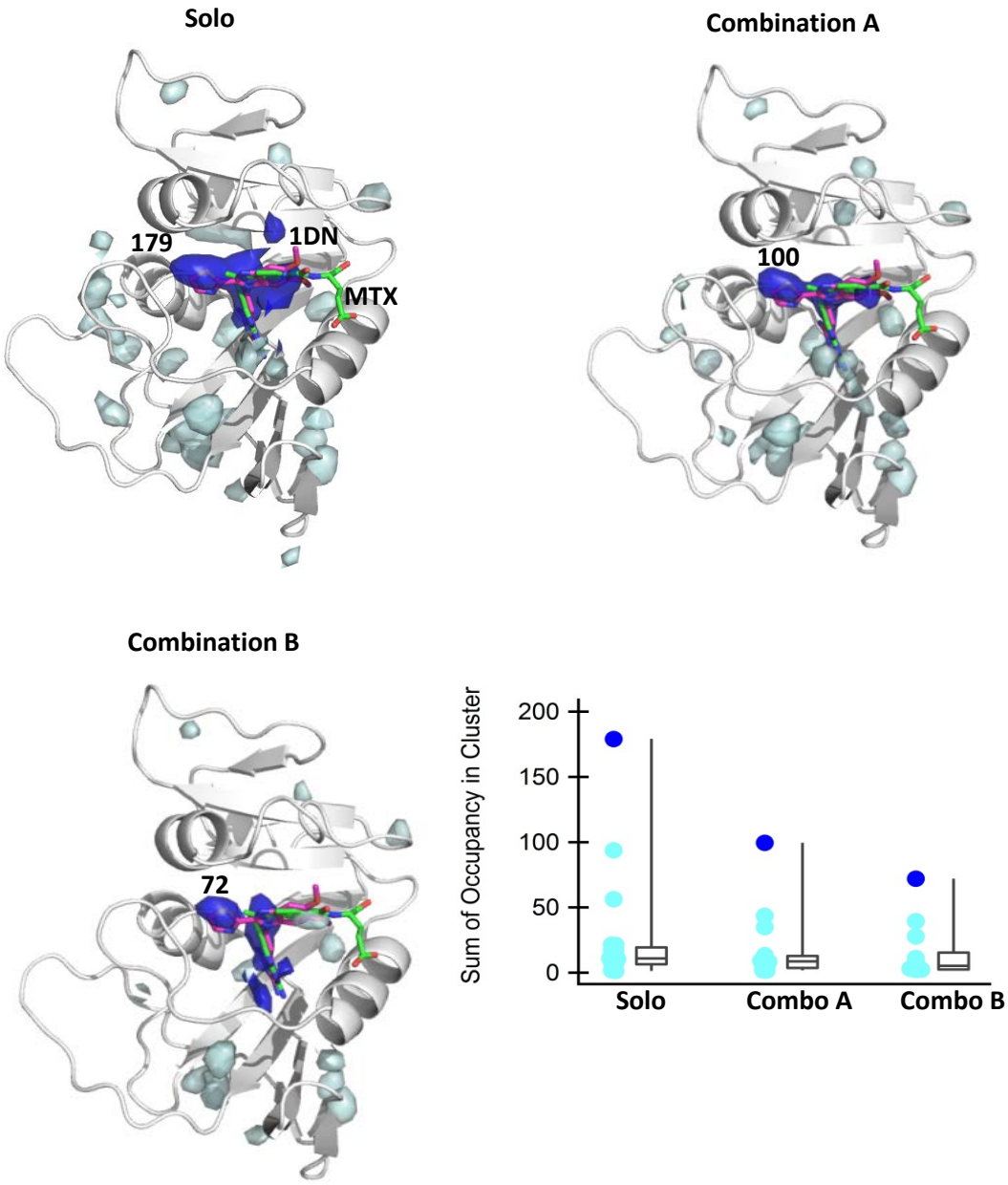


Figure 6.7: The active site of DHFR is correctly identified as the top-ranked site (shown in dark blue) across all three sets of MixMD simulations. The total occupancy for the top sites is given in bold, with the remaining clusters shown in light blue. Methotrexate and the ligand 1DN are shown for reference (PDB:1DF7, MTX and PDB:4LEK,1DN)[176, 264].

Comparing Local Maxima across Solvent Types

As demonstrated in the preceding sections, binding sites can be identified for any of the tested solvent mixtures by considering the total occupancy within a region as mapped by all of the neutral probes. In addition to binding site prediction, however, cosolvent simulations are also frequently used to identify specific interactions of individual probes for use in structure-based drug design. It is possible that solvents run in combination may compete with each other for binding, leading to fewer local sites being identified when solvent mixtures are used rather than solo cosolvent simulations. It is also possible that there may be cooperativity between probes, leading to additional local maxima in adjacent regions that cannot be observed in solo runs.

In order to compare the occupancies across the three sets of simulations, grid points were clustered using the mean shift algorithm implemented in MixMD Probeview to identify local maxima and surrounding points. Comparing the local maxima of each solvent within the active-site region shows differences for some systems between simulations done with each probe individually and those of combined solvent mixtures. For example, simulations of individual probes with ABL kinase identify local maxima for acetonitrile, imidazole, and isopropyl alcohol within the ATP binding portion (left side) of the active site (**Figure 6.8**). In solvent combination B, these three solvents are run in combination. In this case, acetonitrile and imidazole preferentially occupy this site over isopropyl alcohol. This result does not appear to be an artifact of system setup, as none of the simulations (either solo or combined) were initiated with these probe molecules directly in the active site. Moreover, the occupancies shown were generated by averaging over ten individual runs, each with different initial velocities set from a random number seed. Therefore, it appears that the differences in observed occupancies at this site are due to a preference for acetonitrile and imidazole over isopropyl alcohol. While acetonitrile and imidazole still capture the tendency for hydrophobic and aromatic interactions within this region, hydrogen bonding information that may have been captured by isopropyl alcohol is lost. The observed preferential binding also has implications for calculating binding affinities based on probe occupancy. Most cosolvent

methods use the Boltzmann relationship (**Eq. 1**) to calculate binding affinities based on the occupancy of probe molecules. In the event of preferential binding by some probes for a specific site, the non-favored probes would have artificially low occupancies relative to the expected distribution, leading to errors in the calculated binding affinities. Individual probe occupancy for every system is included in the supplementary information.

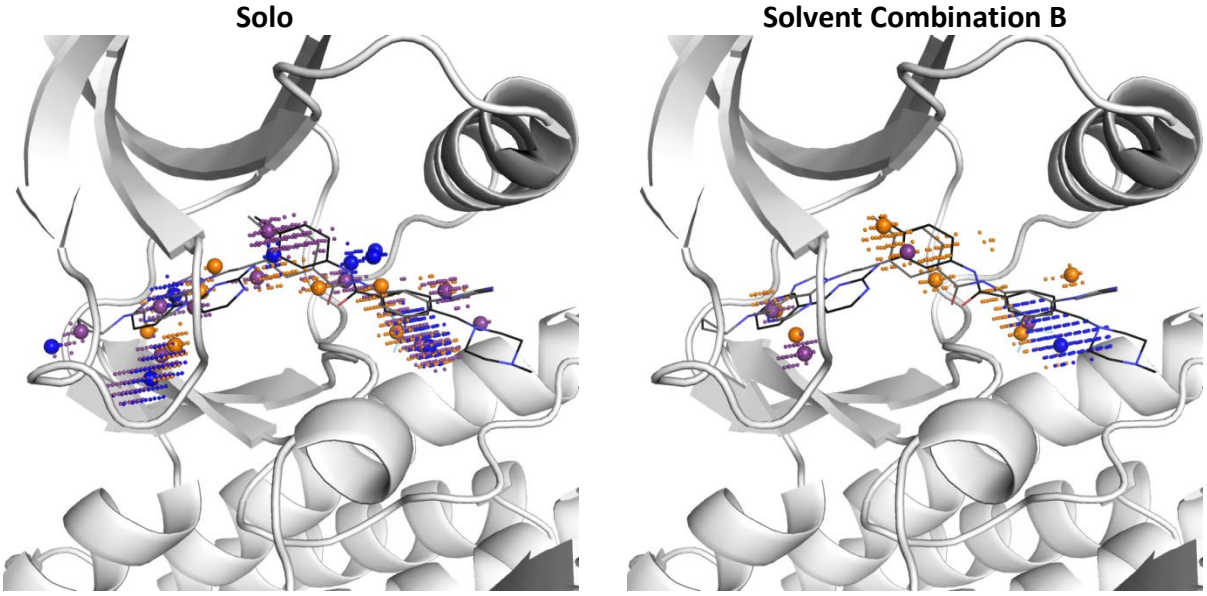


Figure 6.8: Acetonitrile (orange), imidazole (purple), and isopropyl alcohol (blue) grid points with greater than 10% occupancy are shown for the active-site region of ABL kinase. Local maxima are shown as spheres, with surrounding grid points shown. Imatinib (PDB:1OPJ)[250] and B91 (PDB:3KFA)[235] are shown for reference. In the solo simulations, acetonitrile, imidazole, and isopropyl alcohol were each run individually. In the combined set B simulations, these three solvents were run in combination. Relative to the solo simulations, the occupancy in the combined simulations identifies fewer local maxima. For example, the isopropyl occupancy seen in the left portion of the ABL active site is absent in the combined solvent simulations, and it is replaced by imidazole and acetonitrile occupancy.

6.5 Conclusions

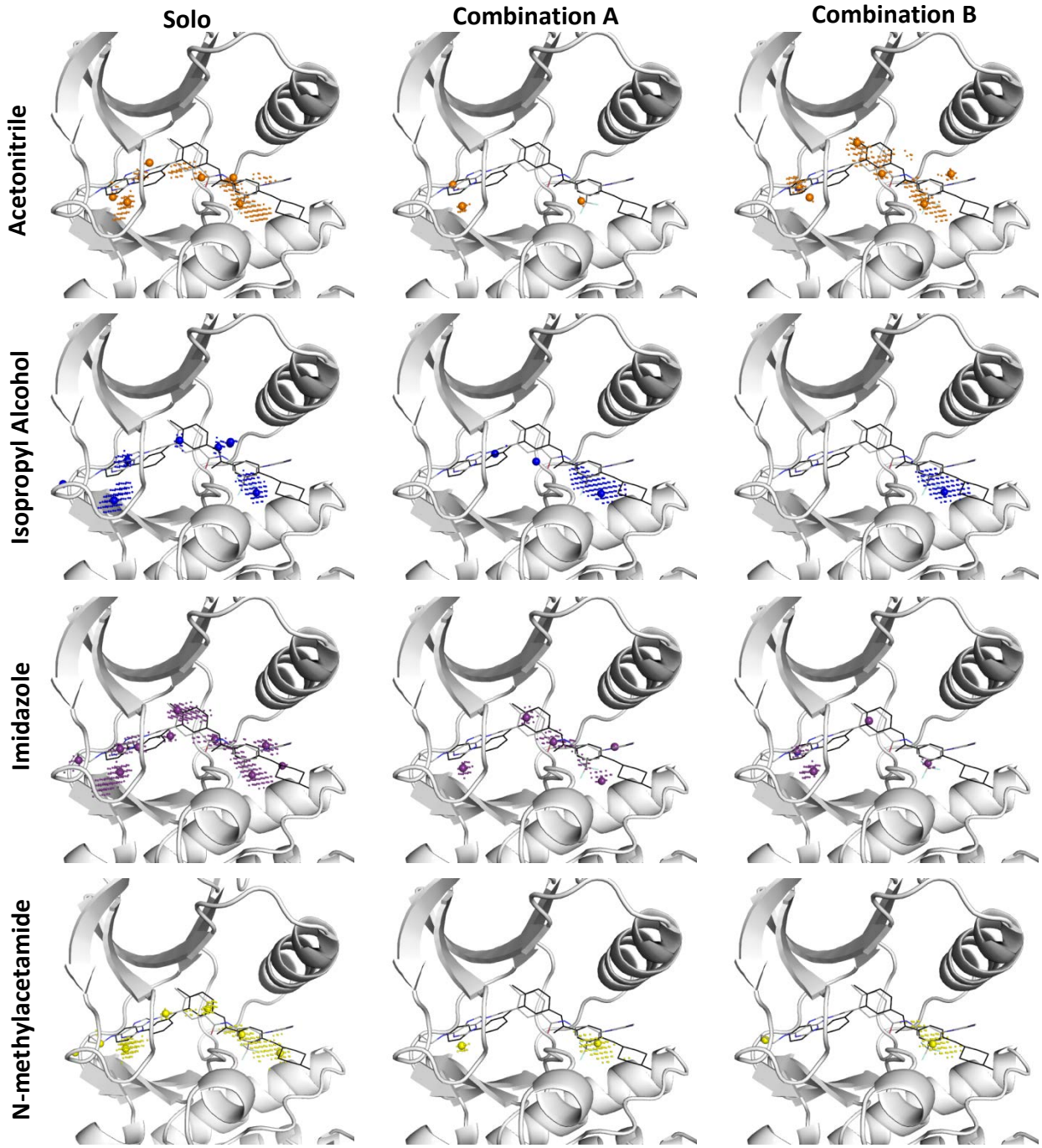
MixMD Probeview successfully identified known active and allosteric sites based on total occupancy of all neutral probe solvents for all systems tested. For each system, the top-ranked site was either the active or allosteric site. For systems having both active and allosteric sites, all of the additional known binding sites were ranked above the remaining sites, with the exception of one set of simulations for AR. As an easy-to-use plugin for the popular

visualization software PyMOL, we expect that MixMD Probeview will facilitate identification of binding sites from cosolvent simulations performed on a wide range of systems.

In addition to Probeview's ability to find regions containing multiple probe molecules, it automates identification and ranking of local maxima of each individual probe solvent by occupancy. Validation studies across both single and combined cosolvent mixtures allowed us to compare the differences in probe sampling across setup procedures. While the top-ranked sites identify the allosteric and active sites for every setup procedure tested, the solo probe simulations show the greatest separation between real binding sites and the rest of the protein surface. As shown in the boxplots in **Figures 6.4-6.7**, when a greater number of simulations are used for analysis, there is a greater separation in total occupancy between known binding sites and less meaningful sites on the protein surface. However, the number of simulations that can be completed is limited by system size and computational resources. Researchers have frequently turned to combined solvent mixtures to reduce the overall number of simulations that must be completed, which appears to be an acceptable choice when the end goal is binding site identification. In regards to mapping all potential interactions within a binding site, the single probe simulations show the best ability to identify all potential interactions. When combined simulations are used, not all local maxima found in solo probe simulations are seen. This is due to other probe types binding more favorably and displacing the other potential probes. Therefore, when the goal of cosolvent simulations is to uncover all potential interactions within a binding site, using single probe solvents appears to be the most reliable choice.

6.6 Supplementary Information

Local Maxima for Simulations of ABL Kinase



Local Maxima for ABL Kinase, continued

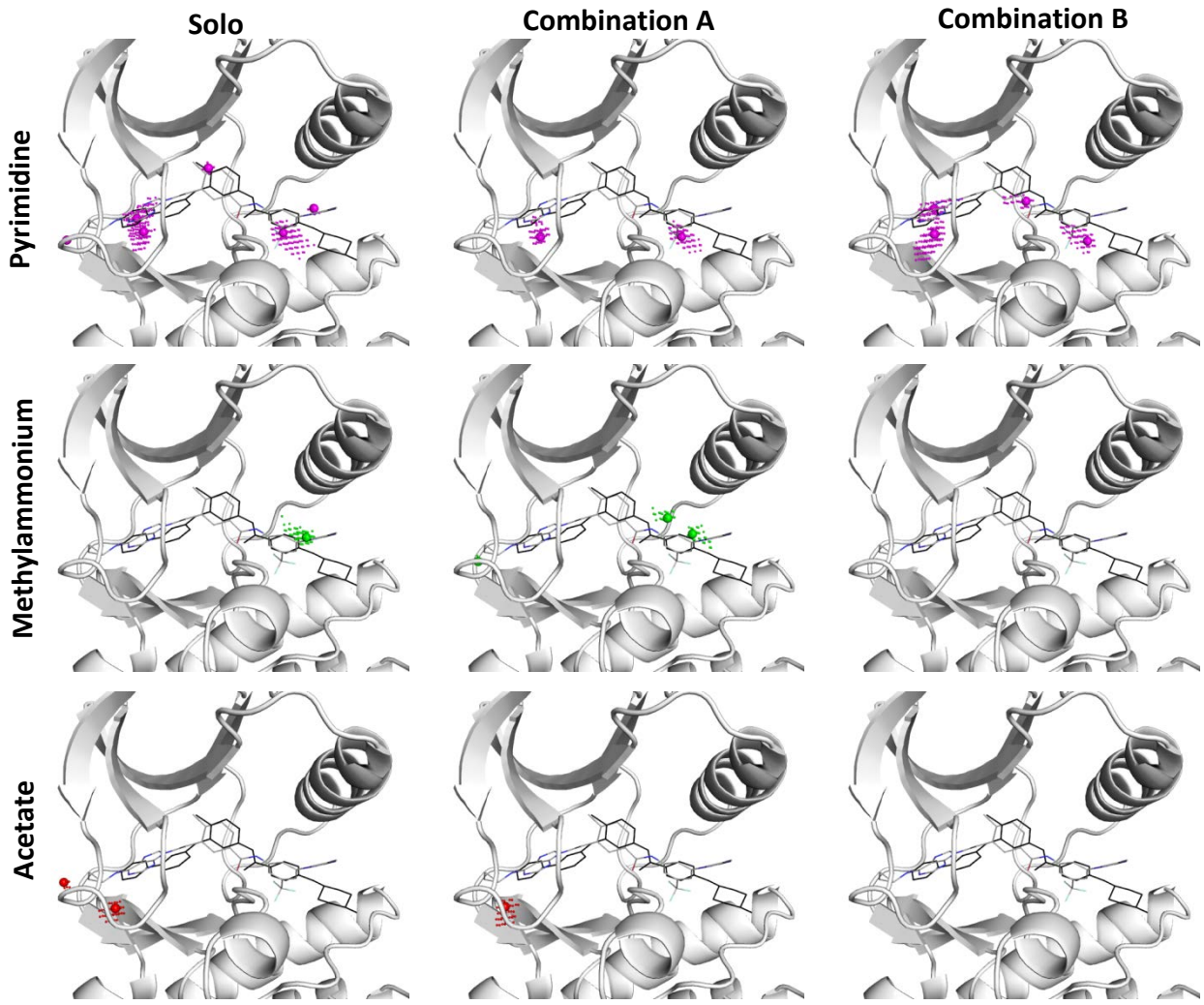
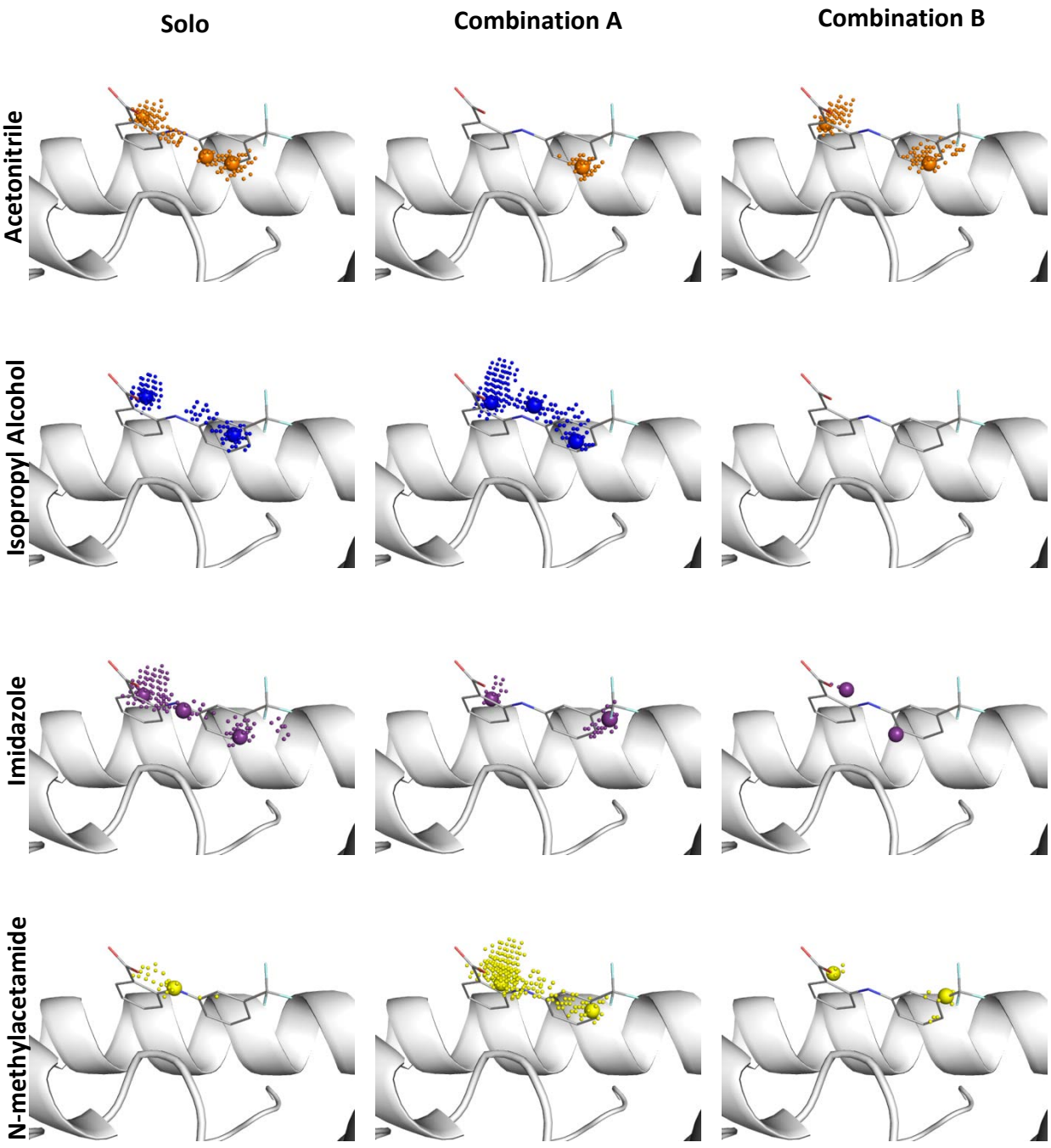


Figure 6.9: MixMD Probeview identified the active site as one of the highest ranked hotspots in ABL kinase. Grid points with 10% or greater occupancy within the active site are shown for each solvent across the three MixMD setups. Local maxima are shown as spheres, with surrounding grid points shown. Imatinib (PDB:1OPJ)[250] and B91 (PDB:3KFA)[235] are shown for reference. Solo simulations accurately map the active site region, in agreement with known ligands. Imidazole shows the most extensive mapping, with local maxima corresponding to aromatic portions of the ligands. Solvent combinations A and B map the active site as well, but with fewer local maxima due to competition between solvents. For example, in solvent combination B the N-methylacetamide occupancy seen within the left-hand side of the ligand in the solo simulations is displaced by pyrimidine. This is consistent with ligand-bound structures which place aromatic rings at this site. However, N-methylacetamide serves to identify hydrogen-bonding interactions, which may not be observed if the site is preferentially bound by other probe molecules.

Local Maxima of Androgen Receptor



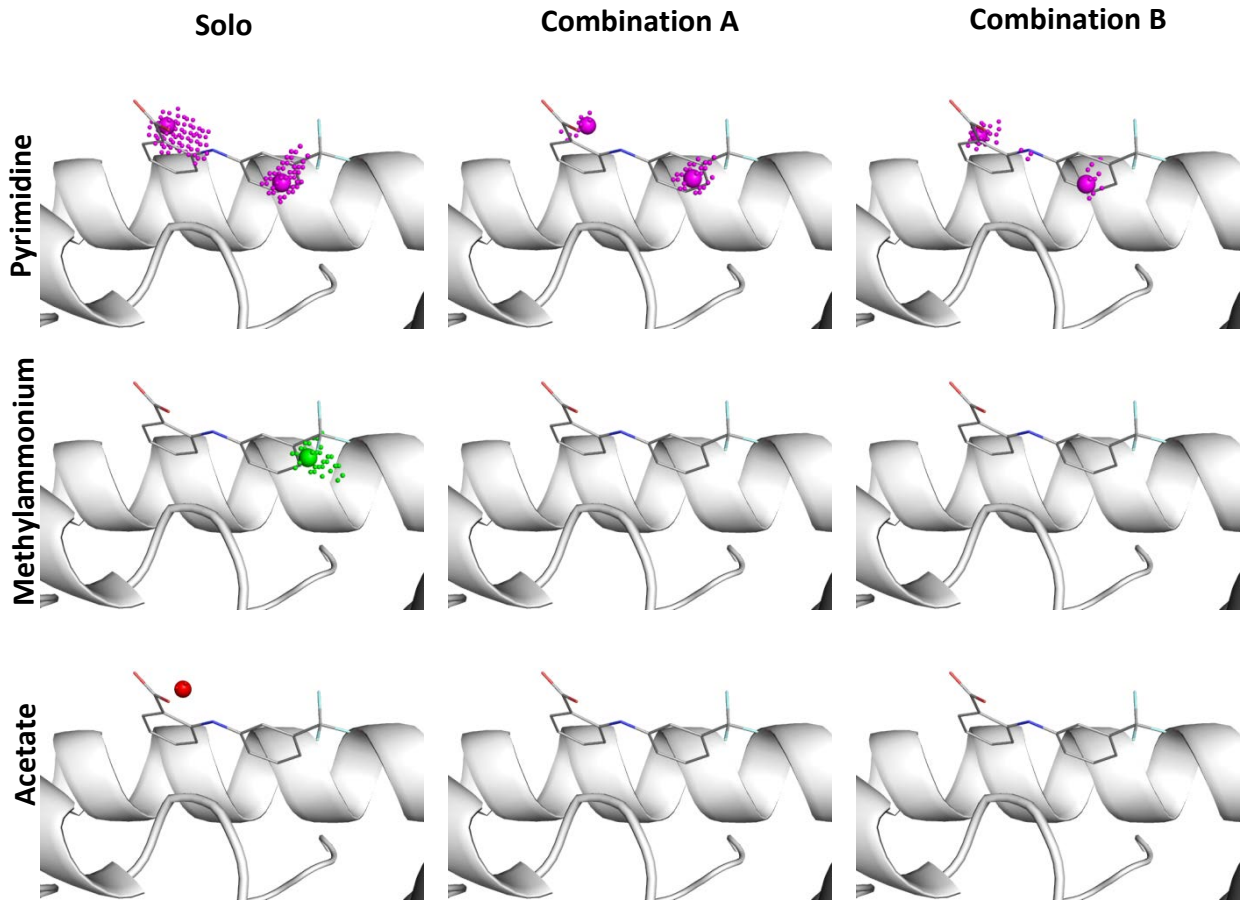
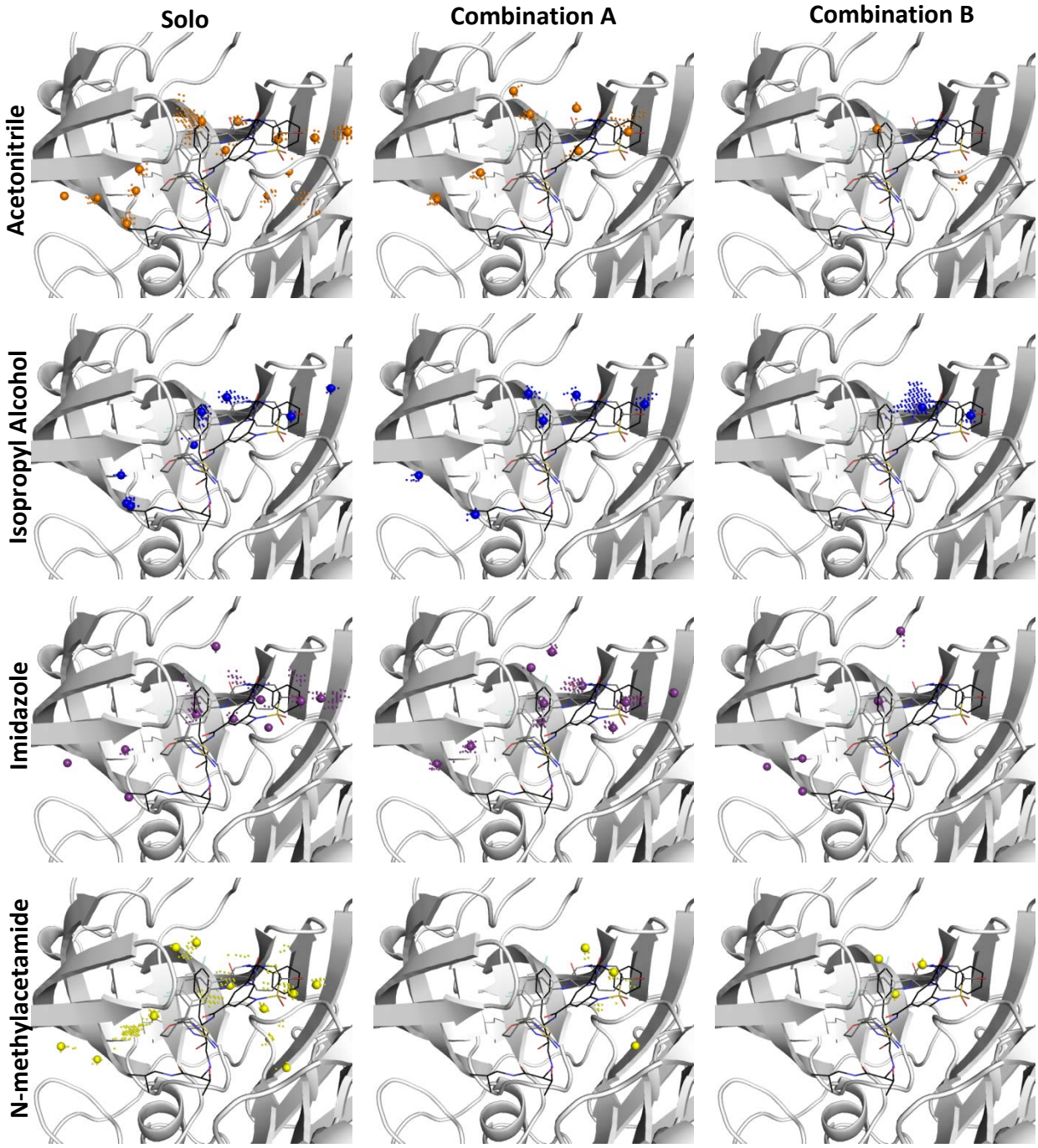


Figure 6.10: MixMD Probeview identified the allosteric site as one of the highest ranked hotspots in androgen receptor. Grid points with 10% or greater occupancy within this site are shown for each solvent across the three MixMD setups. Local maxima are shown as spheres with surrounding grid points shown. The active site of AR has minimal solvent exposure, and so differences in sampling between solvent sets are expected. For this reason, we have shown local maxima for one of the allosteric sites. The allosteric site ligand, flufenamic acid (PDB:2PIX)[254], is shown for reference. Solo simulations show each probe accurately maps the allosteric site ligand but with different occupancy strengths. Acetonitrile, isopropyl alcohol, and imidazole all had similar top occupancies for the solo simulations, with the two charged probes, methylammonium and acetate, having the least occupancy. Solvent combinations A and B mirror the solo simulations, but with a few noticeable differences. First, the charged probes fail to map the ligand at all in both solvent combos A and B. This is likely due to the site's preference for other types of interactions, leading to the charged probe's displacement. Isopropyl alcohol shows strong mapping in combination A, whereas in combination B it is displaced by acetonitrile and imidazole. Visualizing the occupancy at lower levels reveals that isopropyl alcohol does sample this site, but is below the 10% cutoff. Additionally, acetonitrile has only one local maximum in solvent combination A, but two in combination B.

Local Maxima of BACE



Local Maxima of BACE, continued

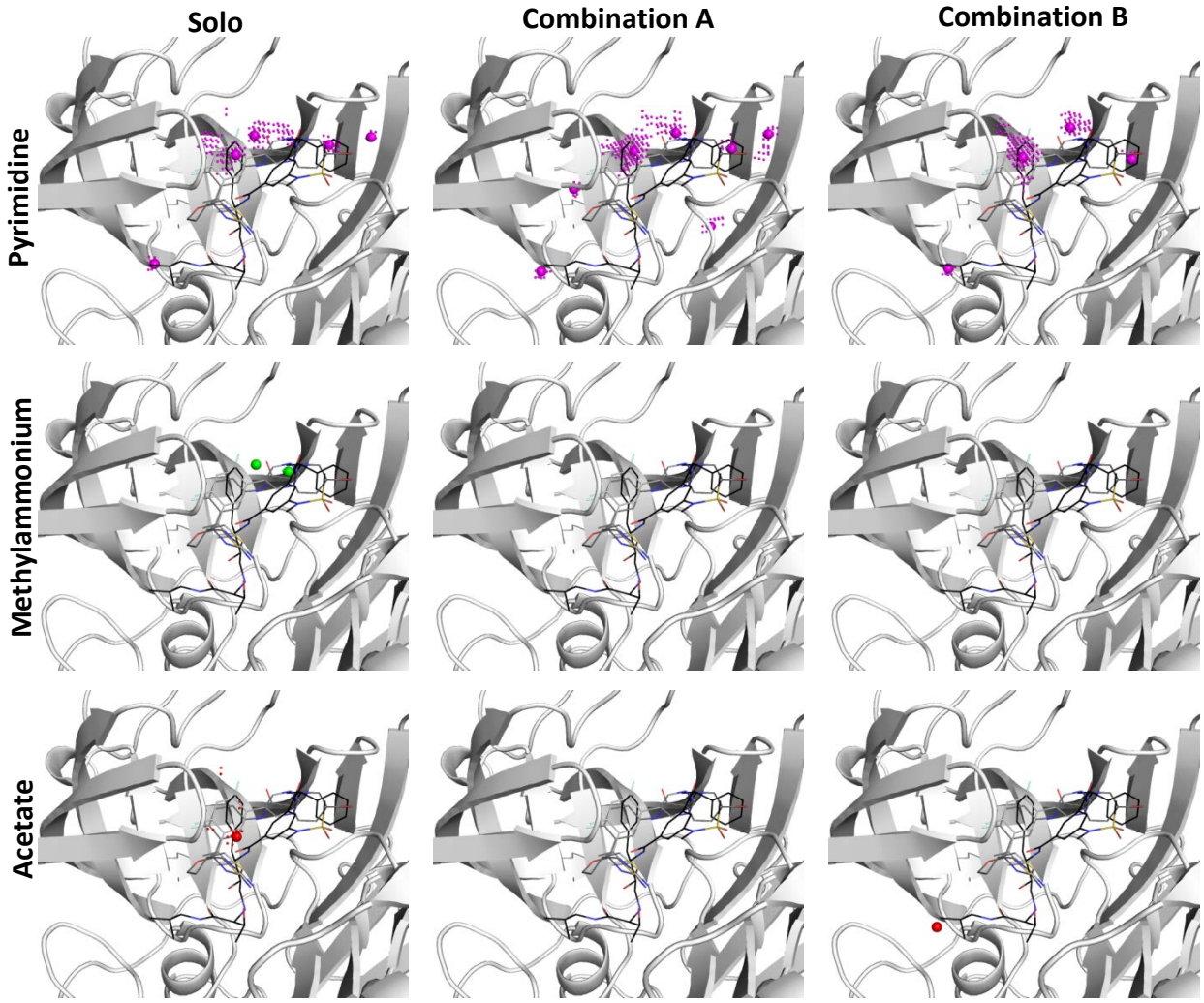
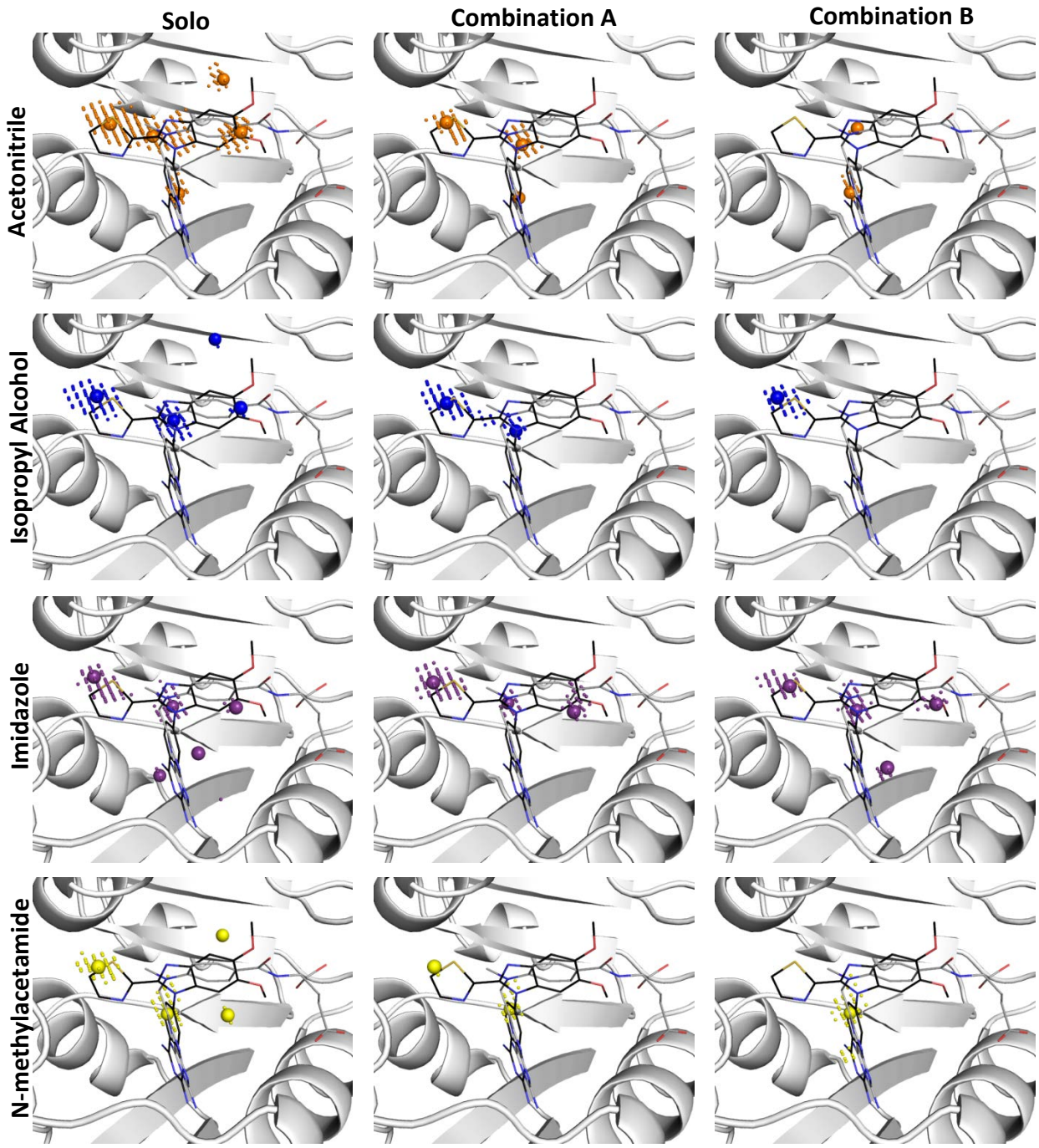


Figure 6.11: MixMD Probeview identified the active site as the highest ranked hotspots in BACE. Grid points with 10% or greater occupancy within the active site are shown for each solvent across the three MixMD setups. Local maxima are shown as spheres, with surrounding grid points shown. Ligands LY2811376 (PDB:4YBI, 4B2)[260], 5E7 (PDB:5DQC)[263], and 7H3 (PDB:5TOL)[262] are shown for reference. Solo simulations show each probe accurately mapping the active site in agreement with known ligands. The neutral probes mapped the active site ligand extensively, while the two charged probes, acetate and methylammonium, had significantly less mapping within the site. Solvent combinations A and B mapped the active site similarly to the solo simulations, with the charged probes being the primary difference. In the combined simulations, the charged probes were displaced in favor of the neutral probes.

Local Maxima of DHFR



Local Maxima of DHFR, continued

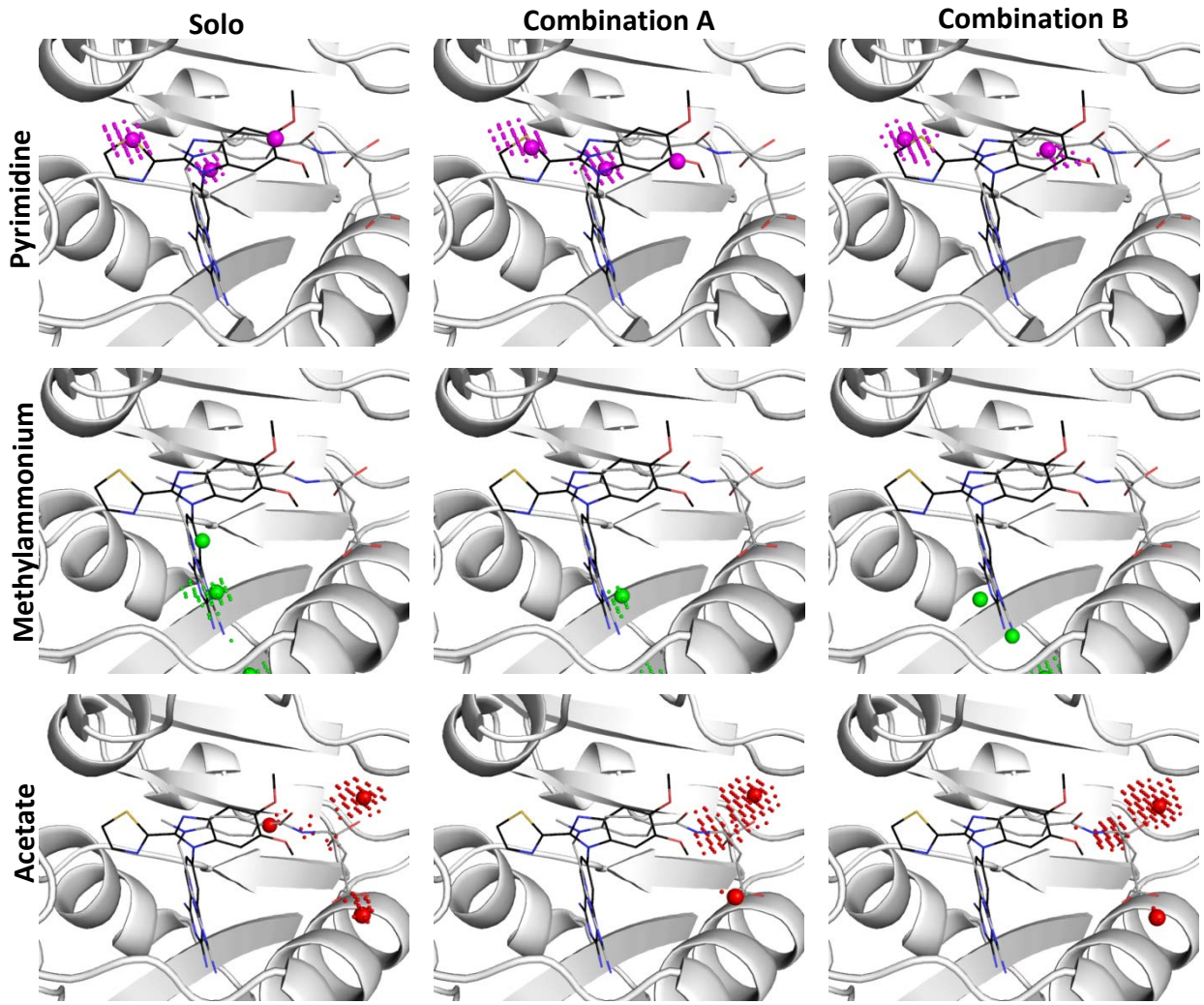


Figure 6.12: MixMD Probeview identified the active site as the highest ranked hotspots in DHFR. Grid points with 10% or greater occupancy within the active site are shown for each solvent across the three MixMD setups, with the exception of the charged probes for which nearby sites are shown. Local maxima are shown as spheres, with surrounding grid points shown. Methotrexate and the ligand 1DN are shown for reference (PDB:1DF7, MTX and PDB:4LEK,1DN)[176, 264]. Mapping of the binding site was similar between all solvents sets, although solvent combination B showed preferential binding to portions of the active-site by acetonitrile and isopropyl alcohol when run in combination with imidazole. The charged probes indicate favorable interactions outside of the core region of the ligand, which mimic the interactions made by the carboxylate groups of methotrexate.

Chapter 7. Conclusions

7.1 Overview and Significant Contributions

The initial chapters of this thesis describe my research contributions aside from development of the mixed-solvent molecular dynamics (MixMD) method. Chapter 2 details the research I completed in fulfillment of the Translational Research Education Certificate (TREC) through the Michigan Institute for Clinical & Health Research (MICHR). This study focused on the spread of extended-spectrum beta-lactamase (ESBL) producing *Escherichia coli,* which are becoming increasingly prevalent. The CTX-M group of ESBLs have become the most common ESBL in *E. coli*[102], and confer resistance to penicillins and 3rd generation cephalosporins[101]. The spread of CTX-M containing *E. coli* is frequently attributed to sequence type 131 (ST131)[107-109]. However, studies of ST131 have primarily focused on resistant isolates, so the contributions of ST131 to the spread of ESBL-positive *E. coli* are difficult to determine. Using a 2006-2008 collection of 1,658 *E. coli* isolates from Gachon University Gil Medical Center in Korea[114], we screened all viable ESBL-positive isolates and a representative sample of ESBL-negative isolates for ST131. We found that among the tested isolates, there was no significant difference in the prevalence of ST131 between ESBL-positive (14% ST131) and -negative (9% ST131) groups[265]. However, ST131 isolates did have greater levels of antibiotic resistance than non-ST131 isolates and were more likely to contain CTX-M-1 groups ESBLs (including CTX-M-15) than other CTX-M types[265]. Additionally, we compared ST131 classification based on two typing methods. The gold-standard for sequence type assignment is multi-locus sequence typing, but this is time-consuming and expensive. Alternatively, the Clermont PCR-based method to identify O25b-ST131[111] or the Weissman method of *fumC* and *fimH* sequencing[112] (CH-typing) may be used to identify ST131-positive isolates. We found that many of the isolates that were classified as ST131 using the PCR-based method were non-ST131 using the sequencing-based CH-typing

method. Furthermore, CH-typing results were in better agreement with previous studies of ST131 prevalence and antibiotic resistance phenotypes. This emphasizes the effect that different testing methods can have on reported prevalence values, which should be taken into consideration when comparing across studies.

Chapter 3 describes our work to better understand the dynamic motions of the H3K36 histone methyltransferase NSD1. The crystal structure of NSD1 has the important post-SET loop in an autoinhibitory position that blocks the entrance of the lysine binding channel[137]. This loop must move in order for the methylation reaction to occur, but it is unclear from the crystal structure if the loop moves spontaneously in solution, or if its movement is induced upon interaction with the nucleosome. A second autoinhibitory conformation is observed in the homologous protein ASH1L[138], which has been shown by mutagenesis studies to be a critical component to enzymatic function[142]. Using long timescale molecular dynamics studies, we simulated the post-SET loop of NSD1 in the crystallographic conformation and the ASH1L conformation. During the simulations, NSD1 adopted three distinct, stable conformations: 1) resembling the starting NSD1 crystal structure, 2) resembling the ASH1L conformation, and 3) rotation of the post-SET loop resulting in widening of the peptide binding cleft[266]. However, in every case the lysine binding channel was still blocked by the auto-inhibitory loop. This suggests that additional interactions, likely with the nucleosome, are required to induce full movement of the autoinhibitory post-SET loop. This finding is consistent with a previous study which combined short timescale MD with docking studies that positioned the nucleosomal DNA against the post-SET loop of NSD1[137].

The remaining chapters and appendices focus on the continued development of the MixMD method. Previous work by our group has primarily focused on validating proper parameters[73] and simulation methods[52]. They have also shown the ability to identify both active and allosteric sites for a number of target proteins[51, 79]. These studies have verified the ability of MixMD to map biologically relevant sites, but do not provide a framework for using MixMD studies in a prospective manner for structure-based drug design. Developments of

other cosolvent simulation techniques, including MDmix and SILCS, have likewise shown the ability to identify binding sites and reproduce favorable interaction sites[63, 72]. A thorough overview of these methods is given in the introduction (Chapter 1).

In the further development of MixMD, we first focused on mapping water within MixMD-identified binding sites (Chapter 4). Cosolvent simulations are performed with a mixture of small molecular probes and water molecules, which directly compete for binding to the protein's surface. Previous studies have focused primarily on the behavior and binding of probe molecules rather than that of water, but MixMD simulations should reproduce favorable water binding sites equally as well as probe binding positions. Moreover, the presence of competing probe molecules allows for the identification of sites which are more favorably occupied by water molecules rather than probes. Using a test set of 10 systems, we performed over 1 μs of MD simulation for the apo structure of each system in the presence of either water or a mixture of water and one of five probe types. This enabled us to determine which water sites were displaceable by specific probe types based on the occupancy level of water at each site. Sites which are highly occupied by water even in the presence of probe molecules are considered to be conserved, while sites which are less frequently occupied by water are considered to be displaceable. Comparison with ligand-bound structures for each of the systems showed a good ability to identify conserved, displaced, and selectively displaced water sites for each system. These occupancy cutoff guidelines can be used to analyze water occupancy from cosolvent simulations, and to determine which sites should be included vs. displaced in structure-based drug design efforts.

Next, we turned our attention to the use of MixMD simulation results for the prospective screening of ligands (Chapter 5). Using the inactive conformation of ABL kinase as a test system, we developed a series of scripts that allow for conversion of MixMD occupancy maps into pharmacophore models for use with the program MOE[30]. First, the AmberTools cpptraj module[157] is used to calculate the occupancy of each functional group within a probe molecule at every point on the protein's surface over the course of the simulation. This yields a

PDB-formatted file of xyz grid points and associated occupancies. The DBSCAN clustering algorithm is used to cluster all grid points above a certain occupancy cutoff into groups. Within each of these clusters, the highest occupancy point is selected to represent the pharmacophore feature's center, and the RMSD of each point within the cluster to the center is calculated to determine the radius. Overlapping clusters are consolidated into joint pharmacophore features when appropriate. Lastly, the pharmacophore model is converted into MOE format to allow for virtual screening. The performance of the pharmacophore model of ABL kinase was tested by screening all ligands that bind to the inactive form of ABL kinase[230-240] and all decoys from the DUD-E ABL inactive set[241] against the pharmacophore model. The stringency of the pharmacophore model was varied by allowing partial matches of the pharmacophore features along with an RMSD multiplier to increase the size of the pharmacophore features. In every model tested, a larger proportion of known active ligands than decoy compounds satisfied the pharmacophore model. Work is currently underway to apply this method in a prospective manner to SRC kinase to screen for new allosteric inhibitors.

Lastly, we sought to simplify the MixMD analysis process in order to make it more accessible to users and facilitate its application to a greater number of systems (Chapter 6). Previously, MixMD analysis was done by loading the resulting probe occupancy maps into PyMOL and manually adjusting the occupancy level to identify high occupancy sites. However, this process is time-consuming and provides a mostly qualitative ranking of occupancy. To aid in the analysis of MixMD data, we developed a PyMOL[100] plugin, which we call MixMD Probeview, to automate the ranking and identification of probe binding sites. MixMD Probeview performs two analyses: first, it serves to identify local maxima from a single probe type and second, it identifies overlapping regions of occupancy from multiple probe types. Identified sites can then be ranked by either maximal or total occupancy. MixMD Probeview was tested on four systems using occupancy results from three different cosolvent simulation procedures. In each case, MixMD Probeview successfully identified known binding sites based on ranking by maximal occupancy.

These developments provide a strong foundation upon which to build future MixMD studies. The MixMD Probeview tool allows for a quantitative comparison of predicted binding sites, allowing for true ligand-binding sites to be distinguished from other easily desolvated regions. The MixMD pharmacophore generation procedure allows for the occupancy maps within these identified regions to be converted into pharmacophore models. This translates the MixMD occupancy maps into a suitable form so that they can be used to inform ligand discovery. Finally, we showed that MixMD simulations can be used to determine a water molecule's potential for displacement. This makes it possible to improve the accuracy of predicted protein-ligand interactions by properly accounting for essential bridging water sites. Altogether, these improvements to the MixMD method are expected to enable its application in a prospective manner, thereby aiding in future drug discovery efforts.

7.2 Future Directions

MixMD simulations are a promising means to identify important interactions on a protein's surface, but they are limited in scope because of the computationally-intensive nature of the simulations. Enhanced sampling methods may be used alongside MixMD simulations to avoid this limitation, by increasing the extent of sampling that can be achieved in a single simulation. This would allow for bigger systems or systems with larger conformational changes to be simulated. Appendix B explores this possibility, by considering the potential of the accelerated molecular dynamics method of McCammon[20] to aid conformational sampling and promote convergence during the simulations. Using HEWL as a test system, accelerated MixMD was found to cut the required simulation time in half relative to standard MixMD simulations. In addition, aMixMD reduced the number of spurious sites that were identified. When applied to systems with known conformational changes, aMixMD was shown to promote conformational sampling. However, the presence of probe molecules also hindered transitions by stabilizing intermediate states. This prevented the full transition between conformational states, but probe molecules did identify sites that are known to be occupied in ligand-bound

structures. These preliminary studies demonstrate the ability of accelerated MixMD to enhance sampling and decrease required computational time.

Future developments could also focus on extending MixMD occupancy maps to be used directly for virtual screening. The current pharmacophore generation procedure represents the MixMD occupancy maps as spheres. While these models show reasonably good performance in virtual screening, it would be desirable to improve their specificity. One way to achieve this would be to develop a virtual screening program that could use the occupancy maps directly. This would include greater shape information, allowing for a more fine-grained representation of the favorable interactions over that of spheres. It would also be possible to incorporate a scoring mechanism into the virtual screening procedure. In standard pharmacophore-based virtual screening, ligands are classified as either satisfying or not satisfying the pharmacophore model. However, this does not take into account how well a ligand satisfies the model. Since the MixMD occupancy maps describe the strength of the interaction at each point, it is possible to identify ligands that best satisfy the pharmacophore model over those that only marginally match the specific interactions. This would allow for ligands that best match the pharmacophore model to be identified and potentially prioritized for further study.

Appendices

Appendix A: Validation of MixMD Setup and Analysis Procedures

Initial Probe Placement

The first Mixed-Solvent Molecular Dynamics (MixMD) studies relied on pre-equilibrated probe and water boxes at a 50%/50% w/w ratio[52, 78, 79]. However, experimental studies of proteins are not typically performed with such high concentrations of organic solvent, making experimental validation of MixMD studies done at 50% concentration difficult. Switching to a 5%/95% v/v concentration of layered probe molecules and water allowed for identification of the same binding sites as studies done at 50%/50% w/w concentrations[79], while ensuring that solvent concentrations did not exceed experimentally feasible values. In the layered setup procedure, the crystal structure of the protein of interest is surrounded with a layer of small molecule probes, followed a box of water in a 5%/95% v/v ratio (**Figure A.1**). A layered setup was chosen rather than pre-equilibrated boxes as this is the most efficient way of facilitating probe sampling at the protein surface. Pre-equilibrated boxes with low probe concentrations would necessitate long simulation times for a sufficient number of probes to reach the protein's surface, and even longer to allow sufficient sampling. In the layered setup, probes are placed near the protein's surface, with the relatively large quantity of surrounding water molecules effectively competing for binding to the protein's surface during the simulation.

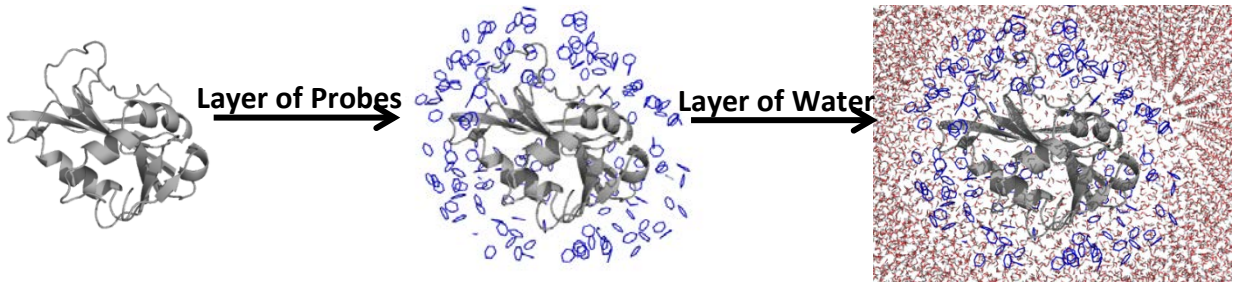


Figure A.1: The current MixMD procedure utilizes a layered cosolvent approach, where the crystal structure of the protein is surrounded with a layer of small molecule probes followed by a box of water molecules.

Nevertheless, it is important to determine the most appropriate method of system setup to ensure that our simulations are effectively and efficiently exploring potential probe binding sites. In the layered approach used in this thesis, the same starting structure is used to initiate each of the ten separate molecular dynamics simulations (ie. the probe and water molecules are in the same locations at the beginning of each simulation). In the first step of a molecular dynamics simulation, each atom in the system must be given an initial random velocity. The current procedure uses a unique random seed number (generated from the current time) for each simulation so that no two simulations are assigned the same initial velocities. This ensures that although each starting structure is the same, the trajectories (and therefore the sampled probe positions) are not identical. However, it is possible that using the same starting structures is biasing our simulations in some manner.

To test this, we performed simulations using Hen Egg White Lysozyme (HEWL) as a test system. MixMD simulations have been previously performed on HEWL and have successfully identified the main binding site as the site having the highest occupancy of probes during the simulation[52]. Using the PACKMOL utility, pyrimidine probe molecules were randomly placed around the structure of HEWL (**Figure A.2**)[267]. Ten separate starting structures were generated. No probe molecules were directly placed in the binding site, with the closest probe molecule positioned more than 4 Å away from the observed acetonitrile binding site (PDB: 2LYO)[77].

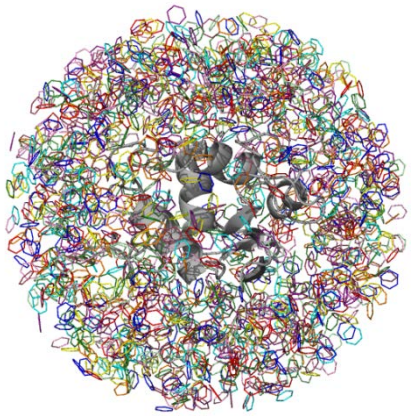


Figure A.2: Starting structures were generated using the PACKMOL utility to randomly place pyrimidine probe molecules around HEWL. Ten such starting structures were generated, each shown in a different color. This setup procedure resulted in varied probe positions, with minimal direct overlap of probe molecules.

Simulations were carried out according to published protocols for 1.75 ns of equilibration and 20 ns of production MD per run[51]. The last 10 ns of each individual trajectory were used for analysis. Following alignment, the trajectories were overlaid with a 0.5 Å cubic grid, and the occupancy of probe molecules at every grid point was calculated separately for each trajectory. To determine if the highest occupancy point successfully identified the active site, the resulting occupancy grids were normalized into units of standard deviations away from the mean (termed σ units), and visualized in PyMOL[100]. In all but 1 of the trajectories, the active site of HEWL is identified as the most occupied site (**Figure A.3**, **Table A.1**). This demonstrates that the active site is identified regardless of initial probe placement. Furthermore, using the same starting structures in each run but with random initial velocities yields the same overall results as using randomly placed probes.

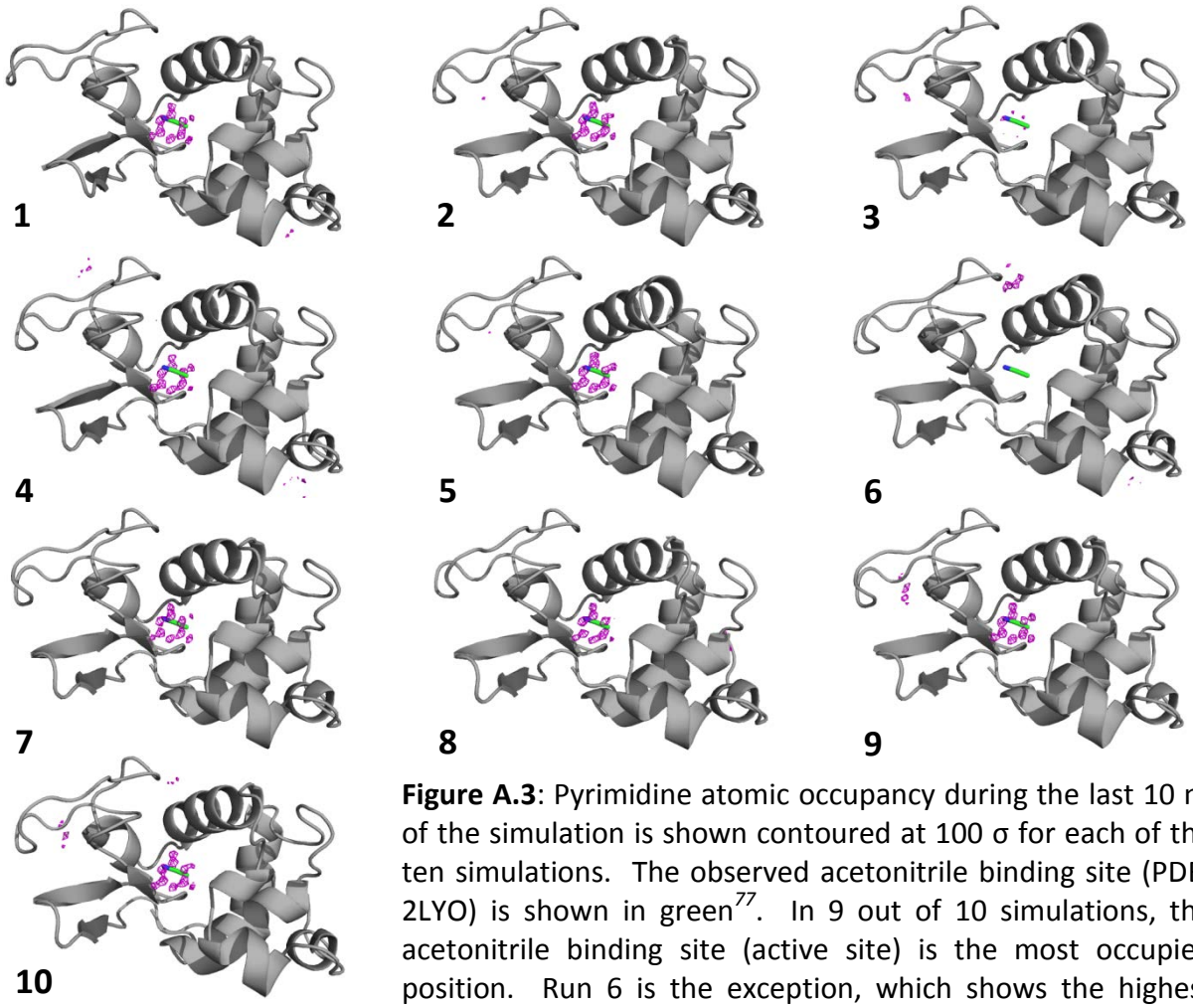


Figure A.3: Pyrimidine atomic occupancy during the last 10 ns of the simulation is shown contoured at 100 σ for each of the ten simulations. The observed acetonitrile binding site (PDB: 2LYO) is shown in green[77]. In 9 out of 10 simulations, the acetonitrile binding site (active site) is the most occupied position. Run 6 is the exception, which shows the highest occupancy site outside of the active-site region.

Run	Highest Probe Occupancy	Location
1	348 σ	active site
2	385 σ	active site
3	174 σ	active site
4	399 σ	active site
5	413 σ	active site
6	170 σ	other
7	390 σ	active site
8	346 σ	active site
9	362 σ	active site
10	350 σ	active site

Table A.1: The highest occupancy and corresponding location in each of the ten simulations is given. The active-site region is indicated by the green acetonitrile in **Figure A.3**.

Adequate Solvent Sampling

A second consideration in the MixMD protocol is the need to ensure adequate sampling. This goal is actually two-fold, in that protein and solvent sampling must both be adequate. Prior work has shown the protein conformational sampling within the standard MixMD protocol to be sufficient, so we turn our focus to the sampling of probe and water positions[52]. Previous MixMD studies have used either the last 5 or 10 ns of a 20 ns production run for analysis of probe occupancy[51, 79]. This allows for the first 10-15 ns to serve as an extended "equilibration", or sampling period, with the goal of minimizing biases from the initial probe placement while allowing for sufficient sampling of potential probe locations. However, this 10-15ns time period was chosen somewhat arbitrarily, and it is not clear if shorter time periods will yield equivalent results. Alternatively, simulation parameters can also be altered to speed up solvent sampling. In order to test for sufficient probe sampling in the context of MixMD, we have compared two sets of simulations. First, standard MixMD simulations of HEWL in pyrimidine and secondly, MixMD simulations of HEWL in pyrimidine with non-hydrogen solvent atoms reassigned a new mass of half the original. These "half-mass" simulations are expected to increase solvent sampling, by allowing for faster solvent motions over the course of the simulation. Ten simulations of HEWL in pyrimidine and water were performed for both the

standard MixMD and "half-mass" setups following standard protocol[51]. The resulting trajectories were aligned, and overlaid with a grid. Segments of 2.5ns from each group of trajectories were then analyzed to identify the position of maximal occupancy in each case (**Figure A.4**). The occupancy is shown as a fraction of the maximal occupancy, ranging from 0 to 1. A value of 1 indicates that the site has the highest occupancy for the tested time period. Values less than 1 are proportional to the maximum value in each segment. The occupancy at the location of the main spurious site in the standard MixMD simulations is also shown for reference on the graph of the "half-mass" simulations.

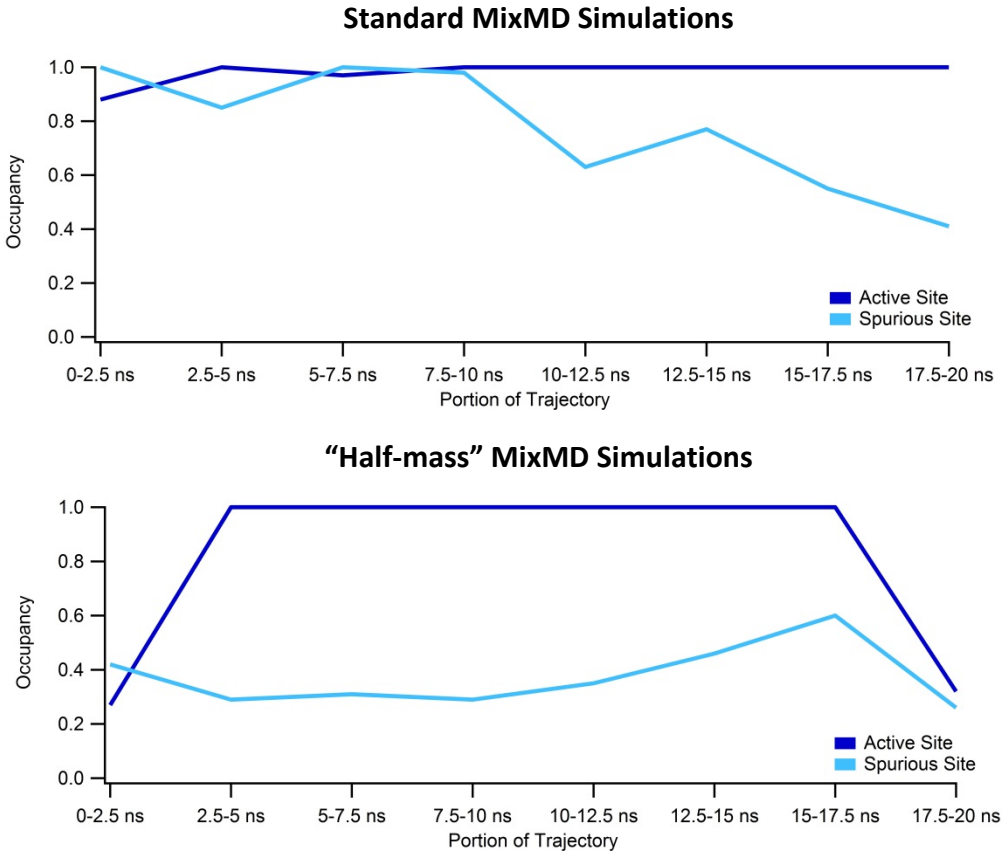


Figure A.4: Total occupancy for pyrimidine across all 10 simulations is shown for each portion of the trajectories. For easier comparison, the occupancy shown is the fraction of the maximum occupancy. For reference, the occupancy of the primary spurious site for the standard MixMD simulations is also shown for the "half-mass" simulations.

Figure A.5 shows the density from the standard MixMD simulations for the first and last 2.5 ns of the 10 trajectories and for the 2.5-5ns portion of the "half-mass" trajectories. Initial portions of the trajectory incorrectly identify a spurious site (circled) as the top-ranked site. After approximately 10 ns (**Figure A.4**), the active site is correctly identified as the top-ranked site in the standard MixMD simulations. Interestingly, the "half-mass" simulations correctly identify the active site of HEWL at earlier time periods. From 2.5 ns onwards, the majority of time periods have the maximum occupancy within the HEWL active-site.

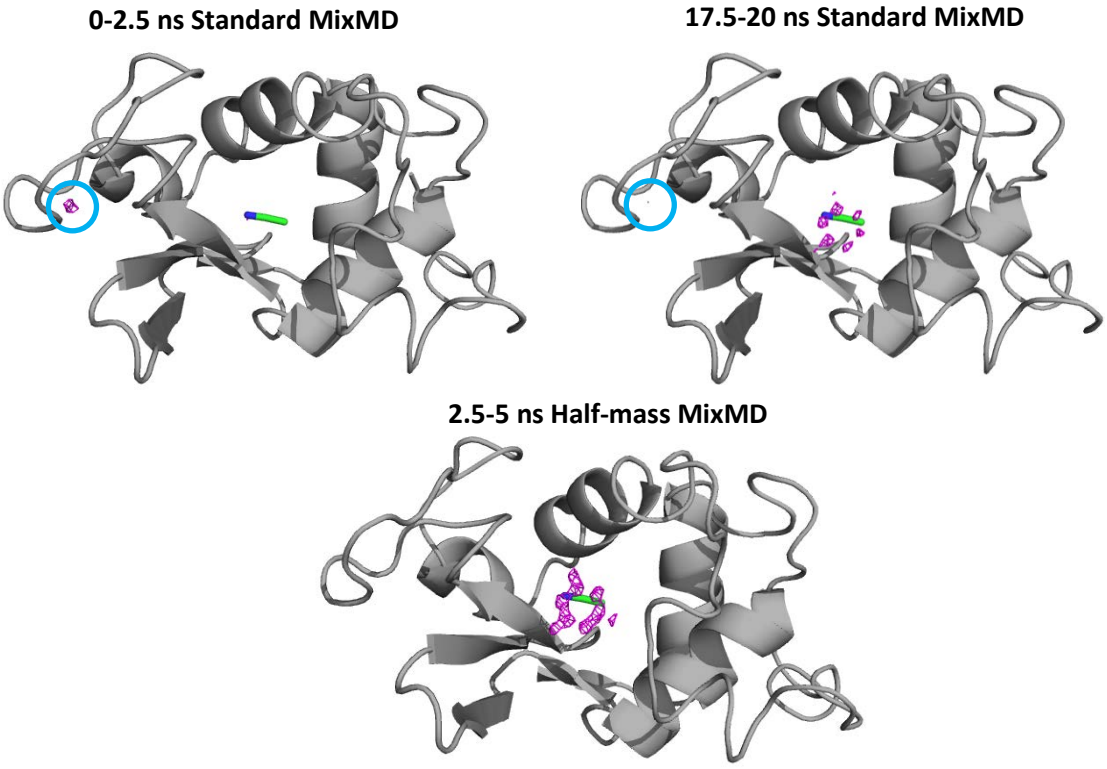


Figure A.5: Left) Pyrimidine atomic occupancy from the first 2.5 ns of all 10 standard MixMD trajectories ranks the spurious site (circled) higher than the active site (Acetonitrile from PDB:2LYO, green stick). Right) Pyrimidine occupancy from the last 2.5 ns of the standard simulations identifies the active site as the top ranked site. Bottom) The 2.5-5 ns time period of the "half-mass" simulations correctly identifies the active site. All figures are contoured at 100 σ.

An important consideration when using altered solvent parameters is the potential to visit unrealistic conformational states. Lighter solvent masses may allow for faster and

potentially larger protein motions than would be observed with standard solvent parameters. In the present simulations, no such behavior was observed, but this may be a possibility with longer simulations or different solvents. As shown in **Figure A.6**, the RMSD of the protein backbone relative to the crystal structure over the course of the production portion of the trajectory is in the 0.5 to 1.5 Å range for both the standard and "half-mass" simulations. Such small values indicate only minor deviation from the starting structure, and are not indicative of unrealistic conformational sampling. Therefore, using "half-mass" solvent parameters are a viable way to increase the efficiency of MixMD simulations. This is possible due to the decrease in required simulation time to achieve sufficiently converged results. However, the standard MixMD analysis procedure using the last 5-10 ns of a 20 ns total trajectory also achieves converged results, without the need for artificially lower solvent parameters.

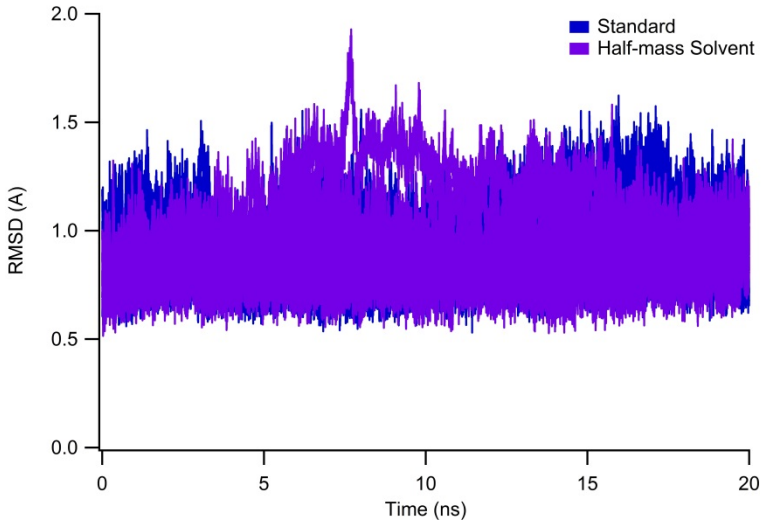


Figure A.6: The backbone RMSD relative to the crystal structure of the production portion of the 10 standard and 10 "half-mass" simulations is shown. Both sets of simulations deviate from the starting structure to a similar extent. RMSD values of 2 Å or less are typically indicative of normal conformational sampling within an MD simulation.

Appendix B: Exploring the Potential of Accelerated Mixed-Solvent Molecular Dynamics Simulations to Enhance Sampling and Capture Conformational Changes

Introduction

Molecular dynamics simulations are routinely used to study conformational changes in proteins, but the extent of such studies are limited by computational resources. In order to generate a sufficient amount of data to yield insight into the dynamics of a protein, a large number of simulations must be completed. This can be extremely difficult, especially in the case of large systems or for systems with conformational changes that occur on long timescales. Enhanced sampling methods, including accelerated molecular dynamics (aMD), are a promising means of bypassing this limitation by increasing the efficiency of conformational sampling[20]. In aMD, the extent of sampling is increased by altering the energy landscape. As shown in **Figure B.1** (reproduced from Figure 1.2 in the introduction), when the potential energy is above a predetermined energy cutoff, the system evolves according to the original energy surface. However, when the energy dips below this cutoff value (E), a modified potential (V(r) + ΔV(r)) is used that decreases the depth of the well:

$$V^*(r) = \begin{cases} V(r), & V(r) \geq E \\ V(r) + \Delta V(r), & V(r) < E \end{cases} \quad (1)$$

By raising the bottom of the well, the energy barrier between states is effectively decreased, leading to accelerated conformational sampling.

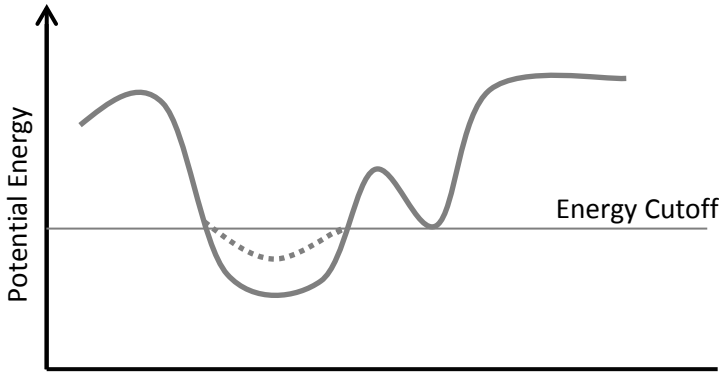


Figure B.1: In the accelerated molecular dynamics method of McCammon[20], a boost is added to the potential energy when the potential energy is below a specified energy cutoff, which effectively decreases the barrier between related conformations. In regions below the energy cutoff, the system evolves according to the modified, "boosted" potential energy surface, depicted as the gray dashed line.

The ability of accelerated dynamics to enhance conformational sampling is especially appealing for structure-based drug design pursuits, which require adequate conformational sampling in order to predict which conformations may be targeted by bound ligands. aMD has previously been used in combination with cosolvent simulation techniques to generate ensembles of protein structures for docking of known small molecule inhibitors to Bcl-2[268]. The ensembles of structures derived from cosolvent aMD simulations resulted in better scores than ensembles from either experimental crystal structures or conventional MD simulations[268]. This increase in performance was attributed to the superior conformational sampling obtained with the combined cosolvent/aMD method. While this study focused on the generation of ensembles of structures for subsequent docking, cosolvent simulations can be used directly to predict binding sites and identify favorable interactions on the protein's surface. MixMD, the cosolvent simulation method developed by our group, has shown an exceptional ability to identify binding sites across a wide range of targets[51]. However, MixMD simulations require a large number of simulations to ensure adequate sampling of probe and water molecules. Furthermore, while some small-scale conformational changes are observed in standard MixMD simulations (such as the bending of the C-terminus in ABL kinase), larger-scale changes such as flipping of the activation loop in ABL kinase are not[51]. In order to test the ability of accelerated

molecular dynamics with MixMD (aMixMD) to enhance conformational sampling, we have performed aMixMD simulations for three systems using multiple levels of boosting.

Methods

Prior to aMD simulations, systems must be heated and equilibrated to the desired temperature. Hen Egg White Lysozyme (HEWL, PDB:2LYO[77]), USP9x deubiquitinase (unpublished crystal structure from Matthew Young, University of Michigan), and ABL Kinase (PDB:1M52[269]) were selected as test systems. Crystallographic waters were removed for HEWL and ABL, but were retained for USP9x deubiquitinase. Hydrogens were added and side chain positions optimized with MolProbity[182] and MOE[30]. The systems were surrounded in a layer of probe molecules (acetonitrile, acetate/methylammonium, isopropyl alcohol, N-methylacetamide, or pyrimidine)[51, 73] followed by a layer of TIP3P water[155] in a 5%/95% v/v ratio. Sodium or chloride ions were added to achieve an overall neutral system. System setup was performed in tleap with the FF99SB force field[247]. Following setup, the systems were minimized for 5000 steps with restraints on the protein and 2500 steps without restraints. The systems were then heated to 300 K at constant volume for 80 ps with a 2 fs time step and restraints of 10 kcal/mol-Å^2 on the protein. Once the correct temperature was reached, the systems were equilibrated for 350 ps at constant pressure while the restraints were gradually removed. A final equilibration step of 1.4 ns without restraints was completed, which was used to calculate the average energy values for the aMD boost levels. Production runs were carried out for a minimum of 20 ns with the GPU enabled version of PMEMD at constant pressure and at 300 K using the Andersen thermostat [185-187]. Accelerated molecular dynamics has been implemented in AMBER, and follows standard simulation procedures with the exception of four additional parameters which control the level of boost applied to the system.

In order to identify appropriate aMD parameters, HEWL was simulated at varying boost levels. HEWL is known to be extremely stable, and so it served as a guide for setting appropriate limits of boosting for MixMD simulations. Boost levels that lead to unfolding in

HEWL are likely to be too great for realistic conformational sampling of other proteins. The dual-boost form of aMD was used,

$$\Delta V(r) = \frac{(Ep - V(r))^2}{\alpha P + (Ep - V(r))} + \frac{(Ed - Vd(r))^2}{\alpha D + (Ed - Vd(r))} \quad (2)$$

where a modified potential ΔV(r) is used to boost both the overall potential and torsional terms[22]. This dual form was introduced to focus the enhanced sampling on the protein while preserving local water structure[22]. The level of boost is controlled by two parameters, a threshold value (E_{thresh}) below which to apply the boost, and a parameter α that controls the strength of the boost. E_{thresh} and α are independently set for the potential and dihedral terms, yielding four parameters which can be adjusted to control the level of accelerated sampling. In order to determine the appropriate levels of boost for use in cosolvent simulations, we tested several different combinations of α_{pot} and $E_{thresh\text{-}pot}$. To avoid excessive boosting to the solvent, we have focused our adjustments on the torsional terms. Across each level of boost, the values of $E_{thresh\text{-}pot}$ and α_{pot} were unchanged, and were given by the following equation:

$$E_{thresh-potential} = Avg.Potential\ Energy + \alpha_{Potential}$$
$$\alpha_{Potential} = N_{atom} * 0.2 \quad (3)$$

where N_{atom} is the total number of atoms in the system. This equation is given in the AMBER manual, and it is suggested to give acceptable performance for most users.

The recommended values for the dihedral threshold and boosting parameter α_D are less clear-cut. The recommended starting dihedral bias α_D is $1/5^{th}$ of the number of residues, multiplied by the approximate energy contribution per degree of freedom (3.5 kcal/mol/residue). In order to account for the presence of probe molecules in our system, we tested several levels of boost, including only protein atoms in the calculation, or both the number of protein residues and the number of probe molecules. In addition, the level of boost

can also be adjusted by adding multiples of α_D to the dihedral energy threshold. This gave us four potential boost levels, given below:

Level 1

$$E_{thresh-D} = Avg.Dihedral\ Energy + 3.5 * N_{RES}$$
$$\alpha_D = 0.2 * 3.5 * N_{RES}$$

Level 2

$$E_{thresh-D} = Avg.Dihedral\ Energy + 3.5 * \left(\frac{N_{RES} + N_{PROBE}}{2}\right)$$
$$\alpha_D = 0.2 * 3.5 * \left(\frac{N_{RES} + N_{PROBE}}{2}\right)$$

Level 3

$$E_{thresh-D} = Avg.Dihedral\ Energy + 3.5 * N_{RES} + 4 * \alpha_D$$
$$\alpha_D = 0.2 * 3.5 * N_{RES}$$

Level 4

$$E_{thresh-D} = Avg.Dihedral\ Energy + 3.5 * (N_{RES} + N_{PROBE}) + 4 * \alpha_D$$
$$\alpha_D = 0.2 * 3.5 * (N_{RES} + N_{PROBE})$$

Higher $E_{thresh-D}$ values will result in higher levels of boost, while higher α_D values result in a modified potential that is closer to the original potential due to their presence in the denominator of Equation 2. Each of these boosting levels was tested on HEWL in order to determine acceptable starting parameters for use on other systems.

Results

HEWL

Ten simulations of 20 ns were completed for each probe type with standard molecular dynamics (no boost) and accelerated MD with each of the four boost levels. In order to assess the stability of HEWL at each level of accelerated sampling, the RMSD relative to the starting crystal structure was calculated over the course of each of the trajectories. A single RMSD summary statistic was generated by averaging over the RMSD vs. time data for each set of

simulations. These values are shown in **Table B.1**. As expected, the standard MD simulations show the smallest mean RMSD values. Typically, simulations which deviate by less than 2 Å relative to the crystal structure are considered to be stable. Accelerated MD simulations using boost levels 1-3 show higher RMSD values compared to the standard MD simulations, but are within the expected range. On the other hand, aMD simulations of HEWL at boost level 4 give mean RMSD values greater than 2 Å, indicating unrealistic conformational sampling. HEWL is an extremely stable protein, containing four disulfide bonds. There are no expected conformational changes in HEWL, and so large RMSD values are indicative of excessive sampling induced by aMD.

Mean RMSD (Å) of HEWL Relative to Crystal Structure over Ten Sets of Simulations

	Pyrimidine	**Acetonitrile**	**Isopropyl Alcohol**	**Acetate/ Methyl-Ammonium**	***N*-methylacetamide**
MD	0.89 ± 0.16	0.87 ± 0.14	0.91 ± 0.18	0.84 ± 0.14	0.87 ± 0.12
aMD 1	1.06 ± 0.23	1.22 ± 0.30	1.19 ± 0.29	1.23 ± 0.35	1.18 ± 0.31
aMD 2	1.09 ± 0.33	1.69 ± 0.56	1.15 ± 0.29	1.81 ± 0.47	1.11 ± 0.27
aMD 3	1.18 ± 0.27	1.97 ± 0.46	1.47 ± 0.38	1.71 ± 0.47	1.33 ± 0.34
aMD 4	1.61 ± 0.38	3.51 ± 0.90	2.74 ± 0.71	3.65 ± 1.12	1.92 ± 0.49

Table B.1: Ten sets of 20 ns simulations were completed for HEWL with either standard or accelerated MixMD with boost levels 1-4. The mean RMSD (Å) ± the standard deviation is shown. These values were calculated by averaging the RMSD relative to the crystal structure over the course of each of the simulations. Only the highest level of boost exceeds the 2 Å limit that is typically used to classify a simulation as stable.

Once appropriate boosting levels have been established, the ability of accelerated molecular dynamics to enhance sampling efficiency can be assessed. For each system and probe type, the current MixMD protocol requires ten simulations of 20 ns, yet only the last 5-10 ns of each simulation are used for analysis. aMD could potentially reduce the length of simulations required by enhancing conformational sampling of the protein and solvent molecules. To test this, we have compared the mapping of the binding site in HEWL between standard MD and accelerated MD simulations. The center-of-mass occupancy over the 5-10ns

portion of the aMixMD boost level 2 simulations and 5-10ns and 10-20ns portion of the standard MixMD simulations was calculated using an in-house modified version of the cpptraj[157] utility in AmberTools[187]. Using the MixMD Probeview tool, high occupancy clusters were identified and ranked based on the total occupancy contained within each cluster. HEWL contains a single ligand binding site at its center. As shown in **Figure B.2**, the active site is identified as the site with the highest total occupancy for each portion of the trajectory. However, the aMixMD simulations showed a much larger difference in occupancy between the top ranked cluster and other spurious sites (**Figure B.3**). For example, the top ranked cluster from the 5-10ns portion of the boost level 2 aMixMD trajectories had an occupancy that was ~19 times greater than the next highest ranked site, compared to a factor of 2.6 between the first and second ranked sites in the 10-20ns portion of the standard MixMD simulations. The same starting structures were used for both standard and accelerated MD runs, with the level of boost being the only difference between simulations. These differences in observed occupancy can therefore be attributed to the additional boost provided by aMD, which promotes more efficient sampling between states and thus likely leads to faster convergence. Boost level 2 was chosen for use in subsequent studies as it allowed for enhanced convergence at a moderate boost level that accounted for both the number of protein residues and the number of probe molecules without excessive sampling. Therefore, while the standard MixMD procedures correctly identify the active site, using aMD with moderate boost levels enhances sampling and facilitates identification of the true ligand binding sites over other spurious sites.

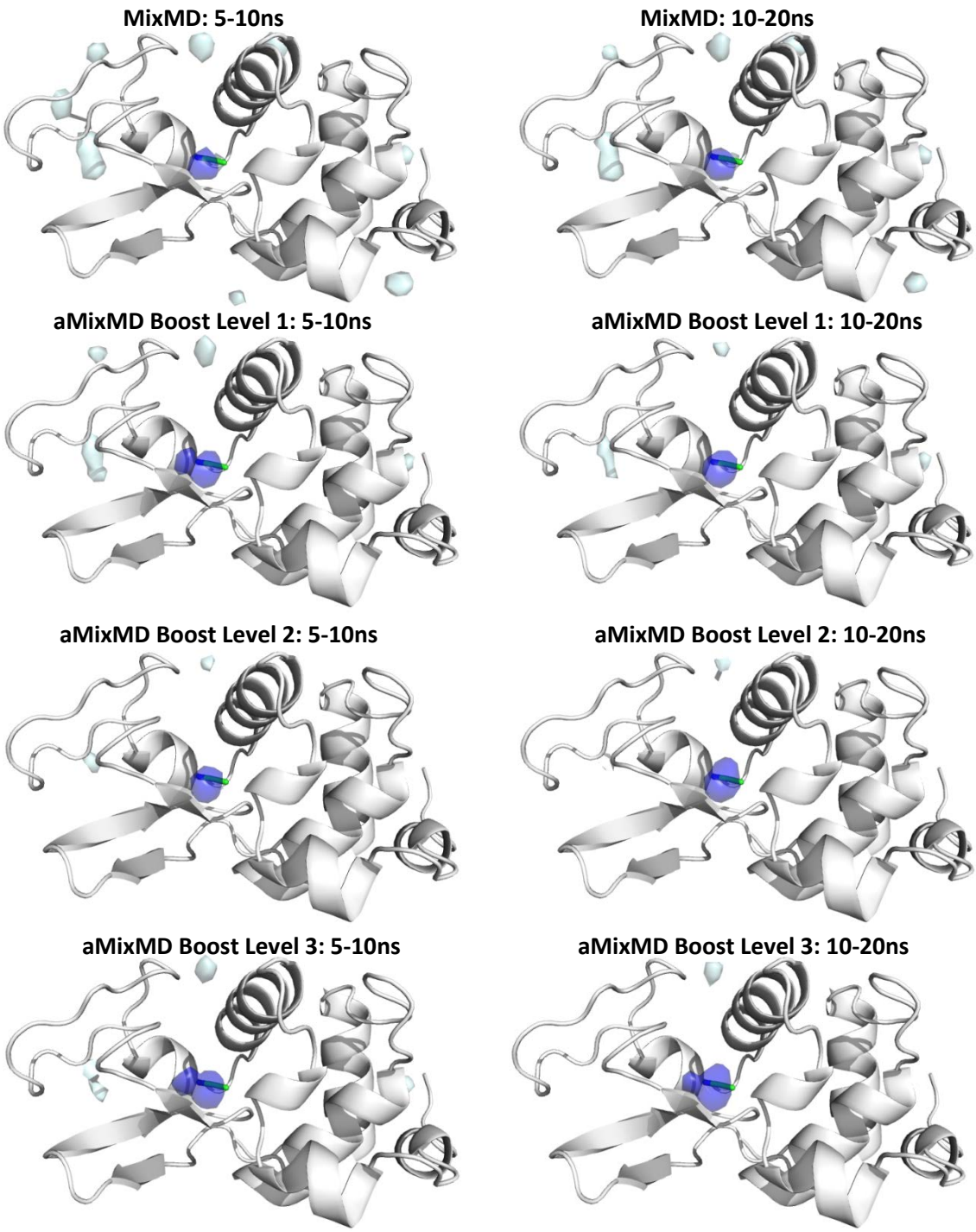


Figure B.2: The highest occupied site identified using the MixMD Probeview tool is shown in dark blue for each of the simulations, with lower occupancy clusters shown in light blue.

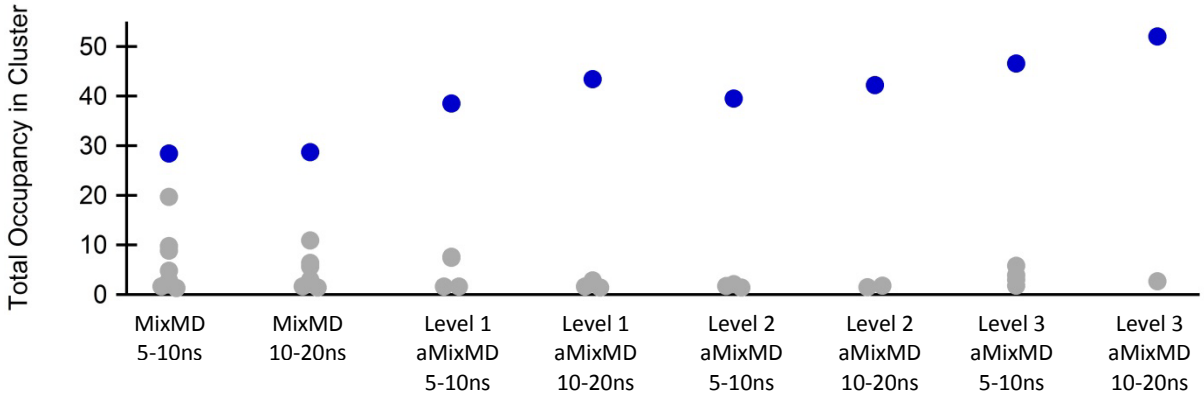


Figure B.3: The graph shows the total occupancy for each cluster in the aMixMD and standard MixMD simulations. The top-ranked site is shown in dark blue, while all other sites are shown in gray. Relative to the standard MixMD simulations, the accelerated MixMD simulations identified fewer spurious sites. As shown in the graph, the difference in total occupancy between the active site and other spurious sites is much larger in the aMixMD simulations, clearly identifying the active site as the top-ranked site.

USP9x Deubiquitinase

Ubiquitination of proteins plays an important role in regulating cellular processes[270]. Mediated by three major classes of enzymes, which have been termed the "readers", "writers", and "erasers", the number and linkage of ubiquitin molecules controls the fate of the target protein[270]. For example, lysine 48 linked ubiquitin chains signal for proteasome-mediated degradation of the protein they are attached to[271]. Deubiquitinases have the opposite effect, by removing ubiquitin molecules from the target protein and thereby blocking subsequent degradation. Within the cell, these enzymes promote a balance between degradation and maintenance of necessary proteins. USP9x is one such deubiquitinase enzyme that is capable of cleaving several types of ubiquitin linkages[272], thereby promoting the survival of the target proteins. Deregulation of USP9x is implicated in a number of disease states, including cancer[272]. For example, MCL1 promotes survival of cells and is overexpressed in several types of cancer. MCL1 would normally be targeted for degradation by the presence of K48 linked ubiquitin chains, but co-occurring overexpression of USP9x upsets this balance, and promotes the maintenance of MCL1[273]. This balance can be restored by knockdown of USP9x in cell lines, which results in decreased levels of MCL1[273]. Inhibition of deubiquitinase with the small molecule WP1130 also results in decreased MCL-1 levels and anti-proliferation of tumor cells[274], demonstrating the potential of deubiquitinases as a therapeutic target[275].

One of the main challenges in targeting deubiquitinases is the need to ensure specificity of the inhibitors. There are almost 100 deubiquitinases which all bind ubiquitin molecules, though with differing specificity for ubiquitin chain types[276]. As each of the deubiquitinases plays a specific role within the cell, it is imperative that inhibitors are sufficiently selective. In order to identify regions on USP9x that may be targeted by inhibitors, we have performed MixMD simulations. Starting from the apo structure of USP9x, ten simulations of 20 ns were completed for each of the probes types with either standard MixMD or accelerated MixMD with boost level 2. Following simulation, the trajectories were aligned and the occupancies over the protein's surface were calculated. MixMD Probeview was used to rank regions based on total probe occupancy over the analyzed time period. As shown in **Figure B.4**, the

simulations identified the active-site region, as well as two main areas of the deubiquitin-ubiquitin interface as binding hotspots. Site 2 for the 5-10ns portion of the aMixMD trajectories is situated near the site of K48 in ubiquitin bound structures. K48 is one of the linkage sites for the formation of ubiquitin chains and K48-linked chains have been shown to be cleavable by USP9x[277]. USP9x has been shown to yield polyubiquitin species (chains of 2+ linked ubiquitins) upon cleavage, indicating that it is capable of binding to and cleaving within the center of a ubiquitin chain rather than just the most distal ubiquitin molecule[278]. Therefore, this site likely corresponds to protein-protein interactions that would occur between ubiquitin chains and USP9x upon binding. The second site, found in both the standard MixMD and accelerated MixMD simulations also occurs at the ubiquitin-deubiquitinase interface. Studies by Ernst and coworkers aimed at producing ubiquitin variants with increased selectivity for specific deubiquitinases found that mutations at this site (among others) were capable of increasing affinity for selected deubiquitinases[279]. As MixMD occupancy is known to identify binding hotspots, this may represent a potential site that could be targeted to block the deubiquitinase-ubiquitin interaction.

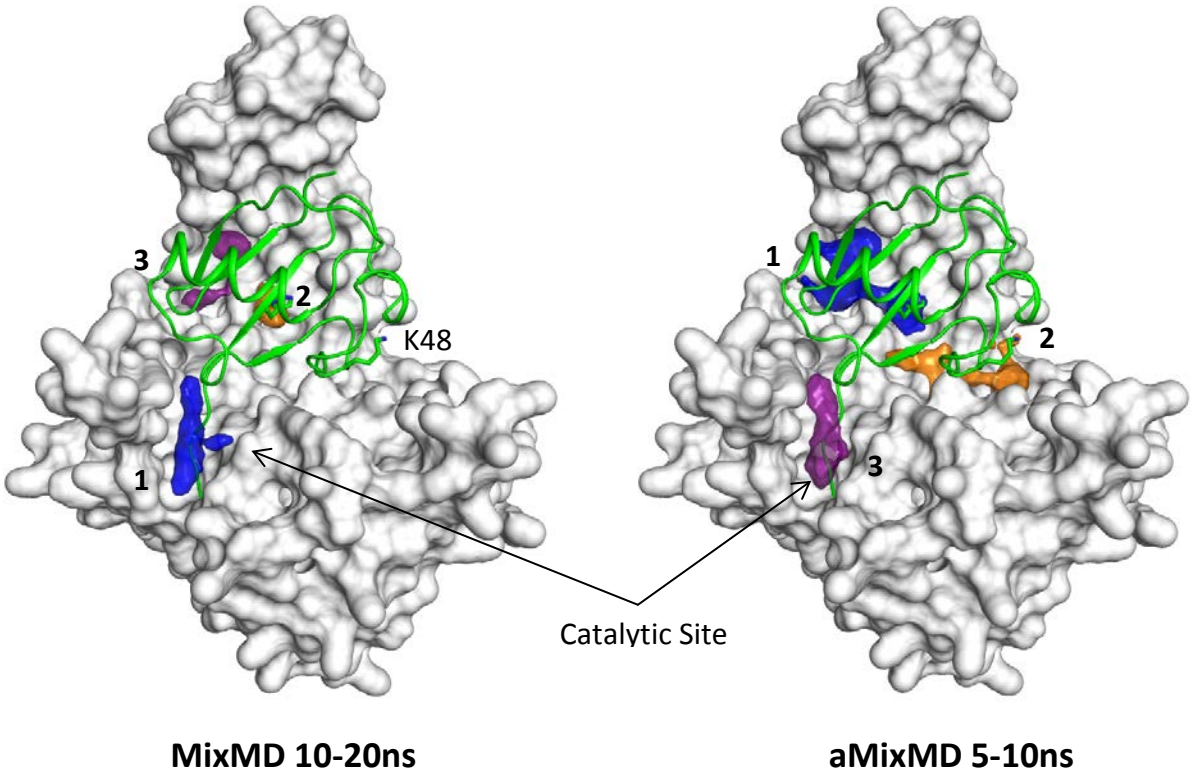


Figure B.4: The top-3 ranked sites by occupancy for the standard MixMD and accelerated MixMD simulations are shown as colored surfaces. Ubiquitin is shown in green for reference, but was not included in the simulations.

ABL Kinase

Kinases undergo a number of conformational changes between active and inactive states[1]. The most well-known and frequently studied conformational change is undoubtedly the DFG-flip. In the active state of kinases, the aspartate of the DFG-motif is oriented into the ATP-binding site ("DFG-in") where it forms important interactions with one of the magnesium ions[1]. In the inactive state, the activation loop reorients, flipping the aspartate away from the active-site and positioning the phenylalanine in its place. Inhibitors targeting the active site region of kinases are classified as type I or type II inhibitors, depending on which conformation they bind to[280]. Type I inhibitors bind to the active conformation, while type II inhibitors bind to the inactive conformation[280]. A number of molecular dynamics studies have been performed on kinases with the goal of capturing this conformational change[281-284]. However, standard molecular dynamics simulations are not sufficient to observe the DFG-flip, leading researchers to utilize enhanced sampling techniques in order to simulate this conformational change. As these different conformations of kinases may be stabilized by inhibitors, enhanced sampling techniques could potentially be combined with cosolvent simulations in order to simultaneously capture conformational changes while determining favorable interactions that may be targeted by inhibitors.

In order to explore this possibility, we have performed aMixMD simulations at multiple boost levels starting from the DFG-out conformation. The proportion of observed DFG-out and DFG-in states is influenced by the protonation state of the aspartate residue of the DFG motif[282]. Protonation of this residue favors the DFG-out conformation while deprotonation favors DFG-in[282]. Therefore, the simulations were performed with a deprotonated Asp which should promote transitions to the DFG-in state. The conformational transition between the DFG-out and DFG-in states was assessed using the metrics developed by Meng et al[281]. As shown in **Figure B.5**, the dihedral angles between the alanine proceeding the DFG-motif and either the aspartate or phenylalanine of the DFG-motif characterize two low-energy states corresponding to the DFG-out and DFG-in conformations.

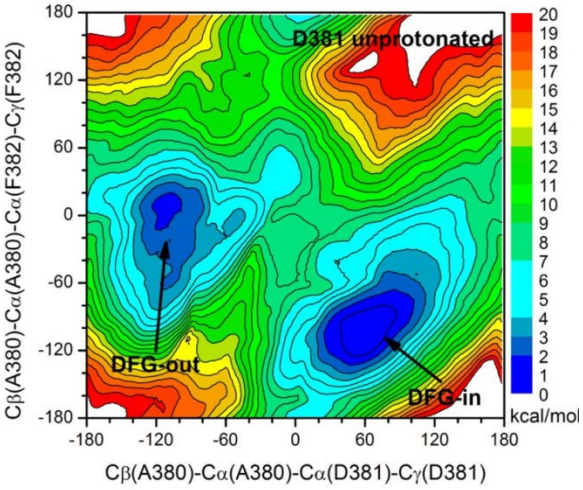


Figure B.5: Adapted from Meng et al[281]. The transition between DFG-in and DFG-out states can be assessed using dihedral angles measured from the preceding alanine to the aspartate of the DFG-motif and from the preceding alanine to the phenylalanine of the DFG-motif.

In order to determine what level of boosting is necessary to induce conformational changes in ABL kinase, several boost levels were tested. Simulations were first performed for ten runs of 20 ns for each of the five solvent types using accelerated MD with boost level 2, yielding 1 μs of total simulation time. This level of boost resulted in conformational sampling that was very close to the initial conformation, as shown in graph A of **Figure B.6**. Performing the same number of simulations using boost level 3 resulted in additional sampling (graph B of **Figure B.6**), but again no transitions to the DFG-in state were observed. aMixMD simulations at boost level 4 caused excessive conformational changes in HEWL, but it is possible that these same boost levels may be suitable to drive sampling in proteins with known or expected conformational changes. Using the original boost level 4, and a modified version, level 4* given below, long timescale simulations of ABL were performed.

Level 4*

$$E_{thresh-D} = Avg.Dihedral\ Energy + 3.5 * (N_{RES} + N_{PROBE})$$

Level 4* results in a smaller amount of boost compared to the original level 4, and was introduced in order to reduce the possibility of unrealistic conformational sampling. Using boost level 4, ten simulations of 40 ns were completed for each of the five solvent types. To account for the lower degree of boost with level 4*, longer simulations of 100 ns were

completed for ten runs with each of the solvent types. As shown in **Figure B.6** panels C and D, both boosting levels resulted in sampling of an intermediate conformational state, with the Asp pointing towards the active site and the phenylalanine positioned upward towards the αC helix. In order to understand the reason for this stabilized intermediate positon, we calculated the occupancy of the probe molecules from all frames in the trajectory having this intermediate conformation (lower right quadrant of **Figure B.6** panel D). The position that is normally occupied by a phenylalanine in the DFG-in state was occupied by probe molecules (panel E of **Figure B.6**), blocking the complete transition to the DFG-in state. This mapping is consistent with known inhibitors of ABL kinase that bind in this pocket, but impedes sampling of the full conformational transition.

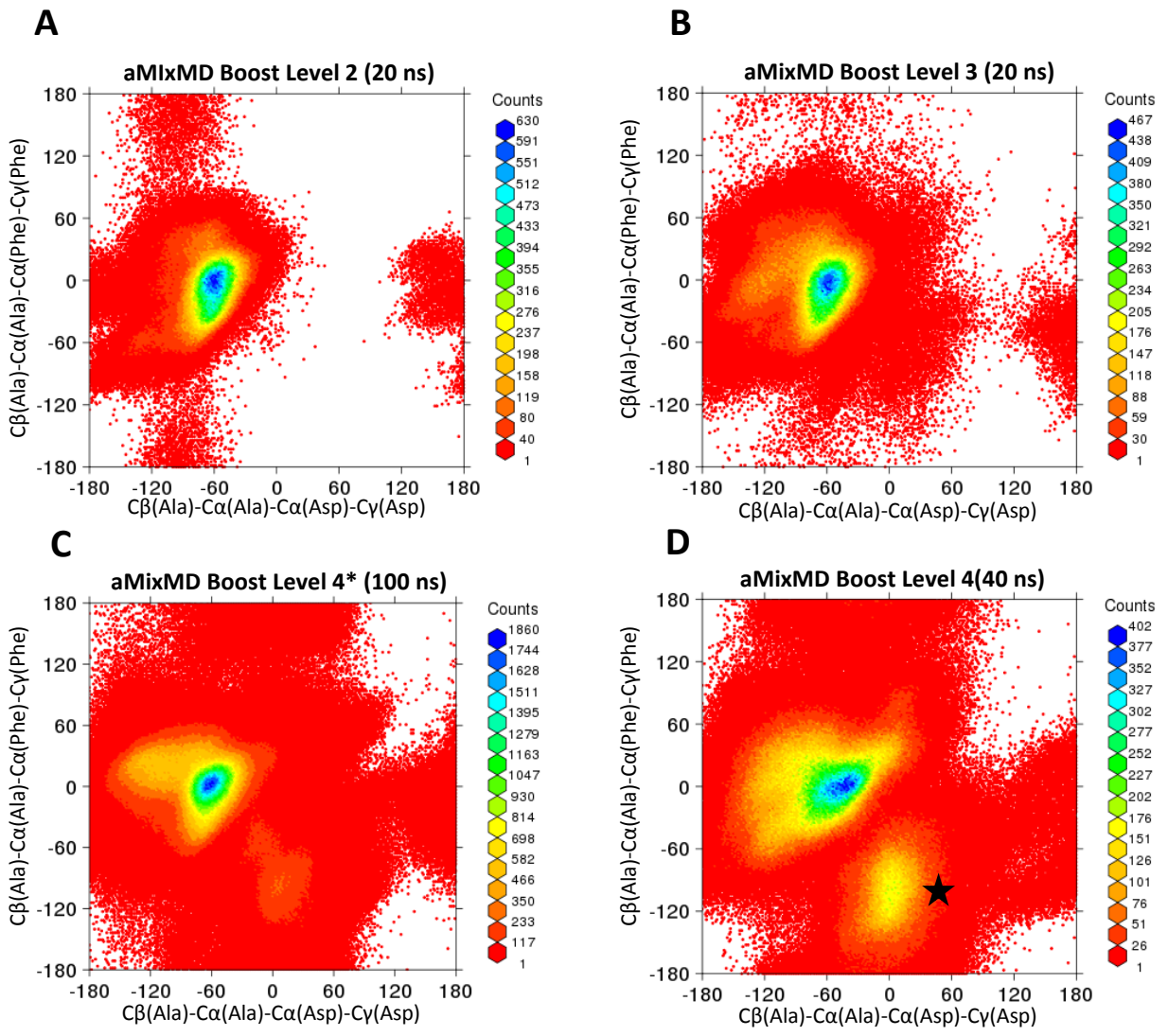


Figure B.6:
A-D) The transition between the DFG-out and DFG-in states is characterized by the Ala-Phe and Ala-Asp dihedral angles. Sampling during the respective trajectories is shown, colored according to the frequency of the observed angles. The black star indicates the dihedral angles characteristic of the DFG-in conformation. E) Pyrimidine occupancy from the frames falling within the frequently sampled region in the right lower quadrant of graph D, from point (-50,-60) to (30,-180). The DFG-out and DFG-in states are shown in green and cyan, respectively. The pyrimidine occupancy overlaps with the region that is occupied by phenylalanine in the DFG-in conformation.

Previous studies have highlighted the ability of cosolvent simulations in combination with enhanced sampling techniques to identify druggable conformations. For example, Kalenkiewicz et al. used aMD, cosolvent simulations, and combined cosolvent aMD simulations to create ensembles of structures that were used for docking[268]. It was found that the combined cosolvent aMD simulations resulted in better docking scores, implying that conformations sampled during the cosolvent aMD simulations were more representative of ligand-bound states than conformations obtained from aMD or cosolvent simulations alone. Studies by Oleinikovas et al. found similar results[285]. Using a combined cosolvent procedure with a Hamiltonian replica exchange-based method of enhanced sampling, they identified cryptic pockets that were not identified with enhanced sampling techniques alone. Our results indicate that while enhanced sampling simulations serve to increase the accessible conformational space within a single simulation, the use of cosolvent probes can also lead to stabilizing new conformations and preventing transitions between states.

Conclusions

These simulations demonstrate the ability of aMD to enhance sampling during MixMD simulations. Low to moderate levels of boost result in increased conformational sampling that is within the expected limits of traditional MD simulations. This additional sampling leads to faster convergence and reduces the number of spurious sites. For systems without known or expected large-scale conformational changes, aMixMD can decrease the computational time required and enable the study of a greater number of systems. On the other hand, higher boosting levels promote more extensive conformational changes, as in the case of ABL kinase. This led to the capture of intermediate conformational states, but it did not allow for full transitions between active and inactive conformations. During simulations of ABL kinase, probe molecules blocked sites typically occupied by side chains in different conformations, which prevented full conformational sampling. It is not clear if this would occur with other systems or if it is a specific effect to ABL kinase. Regardless, aMixMD did promote conformational sampling and transitions between states that are not accessible on normal MD timescales.

References

[1] Taylor, S. S., and Kornev, A. P. (2011) Protein Kinases: Evolution of Dynamic Regulatory Proteins, *Trends in biochemical sciences 36*, 65-77.

[2] Holohan, C., Van Schaeybroeck, S., Longley, D. B., and Johnston, P. G. (2013) Cancer drug resistance: an evolving paradigm, *Nature Reviews Cancer 13*, 714-726.

[3] Davies, J., and Davies, D. (2010) Origins and Evolution of Antibiotic Resistance, *Microbiology and Molecular Biology Reviews 74*, 417-433.

[4] Teague, S. J. (2003) Implications of protein flexibility for drug discovery, *Nature Reviews Drug Discovery 2*, 527-541.

[5] Karplus, M., and McCammon, J. A. (2002) Molecular dynamics simulations of biomolecules, *Nature Structural Biology 9*, 646-652.

[6] Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The Amber biomolecular simulation programs, *Journal of Computational Chemistry 26*, 1668-1688.

[7] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, *Journal of computational chemistry 4*, 187-217.

[8] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005) Scalable molecular dynamics with NAMD, *Journal of computational chemistry 26*, 1781-1802.

[9] Shaw, D.E., Dror, R. O., Salmon, J. K., Grossman, J. P., Mackenzie, K. M., Bank, J. A., Young, C., Deneroff, M. M., Batson, B., Bowers, K. J., Chow, E., Eastwood, M. P., Ierardi, D. J., Klepeis, J. L., Kuskin, J. S., Larson, R. H., Lindorff-Larsen, K., Maragakis, P., Moraes, M. A., Piana, S., Shan, Y., and Towles, B. (2009) Millisecond-Scale Molecular Dynamics Simulations on Anton, *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis,* 1-11.

[10] Sborgi, L., Verma, A., Piana, S., Lindorff-Larsen, K., Cerminara, M., Santiveri, C. M., Shaw, D. E., de Alba, E., and Muñoz, V. (2015) Interaction Networks in Protein Folding via Atomic-Resolution Experiments and Long-Time-Scale Molecular Dynamics Simulations, *Journal of the American Chemical Society 137*, 6506-6516.

[11] Pande, V. S., Baker, I., Chapman, J., Elmer, S. P., Khaliq, S., Larson, S. M., Rhee, Y. M., Shirts, M. R., Snow, C. D., Sorin, E. J., and Zagrovic, B. (2003) Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing, *Biopolymers 68*, 91-109.

[12] Shukla, D., Meng, Y., Roux, B., and Pande, V. S. (2014) Activation pathway of Src kinase reveals intermediate states as targets for drug design, *Nature Communications 5*, 1-11.

[13] Laidler, K. J., and King, M. C. (1983) Development of transition-state theory, *The Journal of Physical Chemistry 87*, 2657-2664.

[14] Sørensen, M. R., and Voter, A. F. (2000) Temperature-accelerated dynamics for simulation of infrequent events, *The Journal of Chemical Physics 112*, 9599-9606.

[15] Abrams, C. F., and Vanden-Eijnden, E. (2010) Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics, *Proceedings of the National Academy of Sciences 107*, 4961-4966.

[16] Sugita, Y., and Okamoto, Y. (1999) Replica-exchange molecular dynamics method for protein folding, *Chemical Physics Letters 314*, 141-151.

[17] Torrie, G. M., and Valleau, J. P. (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling, *Journal of Computational Physics 23*, 187-199.

[18] Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method, *Journal of Computational Chemistry 13*, 1011-1021.

[19] Voter, A. F. (1997) Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events, *Physical Review Letters 78*, 3908-3911.

[20] Hamelberg, D., Mongan, J., and McCammon, J. A. (2004) Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules, *The Journal of Chemical Physics 120*, 11919-11929.

[21] Markwick, P. R., and McCammon, J. A. (2011) Studying functional dynamics in bio-molecules using accelerated molecular dynamics, *Physical Chemistry Chemical Physics 13*, 20053-20065.

[22] Hamelberg, D., de Oliveira, C. A. F., and McCammon, J. A. (2007) Sampling of slow diffusive conformational transitions with accelerated molecular dynamics, *The Journal of Chemical Physics 127*, 155102:1-9.

[23] Miao, Y., Sinko, W., Pierce, L., Bucher, D., Walker, R. C., and McCammon, J. A. (2014) Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation, *Journal of Chemical Theory and Computation 10*, 2677-2689.

[24] Guo, C., and Zhou, H.-X. (2016) Unidirectional allostery in the regulatory subunit RIα facilitates efficient deactivation of protein kinase A, *Proceedings of the National Academy of Sciences 113*, E6776-E6785.

[25] Gasper, P. M., Fuglestad, B., Komives, E. A., Markwick, P. R. L., and McCammon, J. A. (2012) Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities, *Proceedings of the National Academy of Sciences 109*, 21216-21222.

[26] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., and Tropsha, A. (2014) QSAR Modeling: Where have you been? Where are you going to?, *Journal of medicinal chemistry 57*, 4977-5010.

[27] Craig, P. N. (1971) Interdependence between physical parameters and selection of substituent groups for correlation studies, *Journal of medicinal chemistry 14*, 680-684.

[28] Topliss, J. G. (1972) Utilization of operational schemes for analog synthesis in drug design, *Journal of medicinal chemistry 15*, 1006-1011.

[29] Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications, *Nature Reviews Drug Discovery 3*, 935-949.

[30] Molecular Operating Environment (MOE) 2013.08, Chemical Computing Group Inc., Montreal, QC, Canada.

[31] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004) Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, *Journal of medicinal chemistry 47*, 1739-1749.

[32] Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., and Banks, J. L. (2004) Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2.

Enrichment Factors in Database Screening, *Journal of medicinal chemistry 47*, 1750-1759.

[33] Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009) AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility, *Journal of computational chemistry 30*, 2785-2791.

[34] Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982) A geometric approach to macromolecule-ligand interactions, *Journal of molecular biology 161*, 269-288.

[35] Chiem, K., Jani, S., Fuentes, B., Lin, D. L., Rasche, M. E., and Tolmasky, M. E. (2016) Identification of an inhibitor of the aminoglycoside 6'-N-acetyltransferase type Ib [AAC(6')-Ib] by glide molecular docking, *MedChemComm 7*, 184-189.

[36] Yang, S.-Y. (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances, *Drug Discovery Today 15*, 444-450.

[37] Dixon, S. L., Smondyrev, A. M., Knoll, E. H., Rao, S. N., Shaw, D. E., and Friesner, R. A. (2006) PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results, *Journal of computer-aided molecular design 20*, 647-671.

[38] Dixon, S. L., Smondyrev, A. M., and Rao, S. N. (2006) PHASE: a novel approach to pharmacophore modeling and 3D database searching, *Chemical biology & drug design 67*, 370-372.

[39] Wolber, G., and Langer, T. (2005) LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters, *Journal of Chemical Information and Modeling 45*, 160-169.

[40] Waltenberger, B., Garscha, U., Temml, V., Liers, J., Werz, O., Schuster, D., and Stuppner, H. (2016) Discovery of Potent Soluble Epoxide Hydrolase (sEH) Inhibitors by Pharmacophore-Based Virtual Screening, *Journal of Chemical Information and Modeling 56*, 747-762.

[41] Loving, K., Salam, N. K., and Sherman, W. (2009) Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation, *Journal of computer-aided molecular design 23*, 541-554.

[42] Salam, N. K., Nuti, R., and Sherman, W. (2009) Novel Method for Generating Structure-Based Pharmacophores Using Energetic Analysis, *Journal of Chemical Information and Modeling 49*, 2356-2368.

[43] Mattos, C., Bellamacina, C., Peisach, E., Pereira, A., Vitkup, D., Petsko, G., and Ringe, D. (2006) Multiple solvent crystal structures: probing binding sites, plasticity and hydration, *Journal of molecular biology 357*, 1471-1482.

[44] Mattos, C., and Ringe, D. (1996) Locating and characterizing binding sites on proteins, *Nature biotechnology 14*, 595-599.

[45] Allen, K. N., Bellamacina, C. R., Ding, X., Jeffery, C. J., Mattos, C., Petsko, G. A., and Ringe, D. (1996) An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins, *The Journal of Physical Chemistry 100*, 2605-2611.

[46] Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1996) Discovering High-Affinity Ligands for Proteins: SAR by NMR, *Science 274*, 1531-1534.

[47] Goodford, P. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, *Journal of medicinal chemistry 28*, 849-857.

[48] Miranker, A., and Karplus, M. (1991) Functionality maps of binding sites: a multiple copy simultaneous search method, *Proteins 11*, 29-34.

[49] Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., Xia, B., Beglov, D., and Vajda, S. (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins, *Nature protocols 10*, 733-755.

[50] Smith, E. W., Nevins, A. M., Qiao, Z., Liu, Y., Getschman, A. E., Vankayala, S. L., Kemp, M. T., Peterson, F. C., Li, R., Volkman, B. F., and Chen, Y. (2016) Structure-Based Identification of Novel Ligands Targeting Multiple Sites within a Chemokine–G-Protein-Coupled-Receptor Interface, *Journal of medicinal chemistry 59*, 4342-4351.

[51] Ghanakota, P., and Carlson, H. A. (2016) Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems, *The journal of physical chemistry B 120*, 8685-8695.

[52] Lexa, K. W., and Carlson, H. A. (2011) Full Protein Flexibility is Essential for Proper Hot-Spot Mapping, *Journal of the American Chemical Society 133*, 200-202.

[53] Alvarez-Garcia, D., and Barril, X. (2014) Relationship between Protein Flexibility and Binding: Lessons for Structure-Based Drug Design, *Journal of Chemical Theory and Computation 10*, 2608-2614.

[54] Carlson, H. A., and McCammon, J. A. (2000) Accommodating Protein Flexibility in Computational Drug Design, *Molecular Pharmacology 57*, 213-218.

[55] García-Sosa, A. (2013) Hydration properties of ligands and drugs in protein binding sites: tightly-bound, bridging water molecules and their effects and consequences on molecular design strategies, *Journal of chemical information and modeling 53*, 1388-1405.

[56] Breiten, B., Lockett, M., Sherman, W., Fujita, S., Al-Sayah, M., Lange, H., Bowers, C., Heroux, A., Krilov, G., and Whitesides, G. (2013) Water networks contribute to enthalpy/entropy compensation in protein-ligand binding, *Journal of the American Chemical Society 135*, 15579-15584.

[57] Poornima, C., and Dean, P. (1995) Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins, *Journal of computer-aided molecular design 9*, 521-531.

[58] Poornima, C., and Dean, P. (1995) Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions, *Journal of computer-aided molecular design 9*, 500-512.

[59] Poornima, C., and Dean, P. (1995) Hydration in drug design. 2. Influence of local site surface shape on water binding, *Journal of computer-aided molecular design 9*, 513-520.

[60] Raymond, U. L. (1996) How Water Provides the Impetus for Molecular Recognition in Aqueous Solution, *Accounts of Chemical Research 29*, 373-380.

[61] Roberts, B., and Mancera, R. (2008) Ligand-protein docking with water molecules, *Journal of chemical information and modeling 48*, 397-408.

[62] Ghanakota, P., and Carlson, H. A. (2016) Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics, *Journal of medicinal chemistry 59,* 10383-10399.

[63] Guvench, O., and MacKerell, A. (2009) Computational fragment-based binding site identification by ligand competitive saturation, *PLoS computational biology 5*, e1000435.

[64] Raman, E., Yu, W., Guvench, O., and Mackerell, A. (2011) Reproducing crystal binding modes of ligand functional groups using Site-Identification by Ligand Competitive Saturation (SILCS) simulations, *Journal of chemical information and modeling 51*, 877-896.

[65] Lakkaraju, S. K., Raman, E. P., Yu, W., and MacKerell, A. D. (2014) Sampling of Organic Solutes in Aqueous and Heterogeneous Environments Using Oscillating Excess Chemical Potentials in Grand Canonical-like Monte Carlo-Molecular Dynamics Simulations, *Journal of Chemical Theory and Computation 10*, 2281-2290.

[66] Lakkaraju, S. K., Yu, W., Raman, E. P., Hershfeld, A. V., Fang, L., Deshpande, D. A., and MacKerell, A. D. (2015) Mapping Functional Group Free Energy Patterns at Protein Occluded Sites: Nuclear Receptors and G-Protein Coupled Receptors, *Journal of Chemical Information and Modeling 55*, 700-708.

[67] Raman, E., Yu, W., Lakkaraju, S., and Mackerell, A. (2013) Inclusion of Multiple Fragment Types in the Site Identification by Ligand Competitive Saturation (SILCS) Approach, *Journal of chemical information and modeling 53*, 3384-3398.

[68] Raman, E. P., Vanommeslaeghe, K., and MacKerell, A. D. (2012) Site-Specific Fragment Identification Guided by Single-Step Free Energy Perturbation Calculations, *Journal of Chemical Theory and Computation 8*, 3513-3525.

[69] Raman, E. P., Lakkaraju, S. K., Denny, R. A., and MacKerell, A. D., Jr. (2016) Estimation of relative free energies of binding using pre-computed ensembles based on the single-step free energy perturbation and the site-identification by Ligand competitive saturation approaches, *J Comput Chem 38*, 1238-1251.

[70] Yu, W., Lakkaraju, S. K., Raman, E. P., Fang, L., and MacKerell, A. D. (2015) Pharmacophore Modeling Using Site-Identification by Ligand Competitive Saturation (SILCS) with Multiple Probe Molecules, *Journal of Chemical Information and Modeling 55*, 407-420.

[71] Yu, W., Lakkaraju, S. K., Raman, E. P., and MacKerell, A. D. (2014) Site-Identification by Ligand Competitive Saturation (SILCS) assisted pharmacophore modeling, *Journal of computer-aided molecular design 28*, 491-507.

[72] Seco, J., Luque, F., and Barril, X. (2009) Binding site detection and druggability index from first principles, *Journal of medicinal chemistry 52*, 2363-2371.

[73] Lexa, K. W., Goh, G. B., and Carlson, H. A. (2014) Parameter Choice Matters: Validating Probe Parameters for Use in Mixed-Solvent Simulations, *Journal of chemical information and modeling 54*, 2190-2199.

[74] Carlson, H. A., Masukawa, K. M., and McCammon, J. A. (1999) Method for Including the Dynamic Fluctuations of a Protein in Computer-Aided Drug Design, *The Journal of Physical Chemistry A 103*, 10213-10219.

[75] Meagher, K. L., and Carlson, H. A. (2004) Incorporating protein flexibility in structure-based drug discovery: using HIV-1 protease as a test case, *Journal of the American Chemical Society 126*, 13276-13281.

[76] Carlson, H. A., Masukawa, K. M., Rubins, K., Bushman, F. D., Jorgensen, W. L., Lins, R. D., Briggs, J. M., and McCammon, J. A. (2000) Developing a Dynamic Pharmacophore Model for HIV-1 Integrase, *Journal of medicinal chemistry 43*, 2100-2114.

[77] Wang, Z., Zhu, G., Huang, Q., Qian, M., Shao, M., Jia, Y., and Tang, Y. (1998) X-ray studies on cross-linked lysozyme crystals in acetonitrile-water mixture, *Biochimica et biophysica acta 1384*, 335-344.

[78] Lexa, K., and Carlson, H. (2013) Improving protocols for protein mapping through proper comparison to crystallography data, *Journal of chemical information and modeling 53*, 391-402.

[79] Ung, P. M. U., Ghanakota, P., Graham, S. E., Lexa, K. W., and Carlson, H. A. (2016) Identifying binding hot spots on protein surfaces by mixed-solvent molecular dynamics: HIV-1 protease as a test case, *Biopolymers 105*, 21-34.

[80] Nussinov, R., and Tsai, C. J. (2012) The different ways through which specificity works in orthosteric and allosteric drugs, *Current pharmaceutical design 18*, 1311-1316.

[81] Lemmon, G., and Meiler, J. (2013) Towards ligand docking including explicit interface water molecules, *PloS one 8*, e67536.

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

[83] Garcia-Sosa, A. T., Mancera, R. L., and Dean, P. M. (2003) WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes, *Journal of molecular modeling 9*, 172-182.

[84] Barillari, C., Taylor, J., Viner, R., and Essex, J. (2007) Classification of water molecules in protein binding sites, *Journal of the American Chemical Society 129*, 2577-2587.

[85] Amadasi, A., Spyrakis, F., Cozzini, P., Abraham, D., Kellogg, G., and Mozzarelli, A. (2006) 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*, 289-309.

[86] Rossato, G., Ernst, B., Vedani, A., and Smiesko, M. (2011) AcquaAlta: a directional approach to the solvation of ligand-protein complexes, *Journal of chemical information and modeling 51*, 1867-1881.

[87] Michel, J., Tirado-Rives, J., and Jorgensen, W. (2009) Prediction of the water content in protein binding sites, *The journal of physical chemistry B 113*, 13337-13346.

[88] Michel, J., Tirado-Rives, J., and Jorgensen, W. (2009) Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization, *Journal of the American Chemical Society 131*, 15403-15411.

[89] Bayden, A. S., Moustakas, D. T., Joseph-McCarthy, D., and Lamb, M. L. (2015) Evaluating Free Energies of Binding and Conservation of Crystallographic Waters Using SZMAP, *Journal of Chemical Information and Modeling 55*, 1552-1565.

[90] Cui, G., Swails, J. M., and Manas, E. S. (2013) SPAM: A Simple Approach for Profiling Bound Water Molecules, *Journal of Chemical Theory and Computation 9*, 5539-5549.

[91] Lazaridis, T. (1998) Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids, *The Journal of Physical Chemistry B 102*, 3542-3550.

[92] Lazaridis, T. (1998) Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory, *The Journal of Physical Chemistry B 102*, 3531-3541.

[93] Young, T., Abel, R., Kim, B., Berne, B., and Friesner, R. (2007) Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding, *Proceedings of the National Academy of Sciences 104*, 808-813.

[94] Abel, R., Young, T., Farid, R., Berne, B., and Friesner, R. (2008) Role of the active-site solvent in the thermodynamics of factor Xa ligand binding, *Journal of the American Chemical Society 130*, 2817-2831.

[95] Li, Z., and Lazaridis, T. (2011) Computing the thermodynamic contributions of interfacial water, *Methods in molecular biology 819*, 393-404.

[96] Nguyen, C. N., Young, T. K., and Gilson, M. K. (2012) Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril, *The Journal of Chemical Physics 137*, 44101.

[97] Ramsey, S., Nguyen, C., Salomon-Ferrer, R., Walker, R. C., Gilson, M. K., and Kurtzman, T. (2016) Solvation thermodynamic mapping of molecular surfaces in AmberTools: GIST, *Journal of Computational Chemistry 37*, 2029-2037.

[98] Haider, K., and Huggins, D. (2013) Combining solvent thermodynamic profiles with functionality maps of the Hsp90 binding site to predict the displacement of water molecules, *Journal of chemical information and modeling 53*, 2571-2586.

[99] Alvarez-Garcia, D., and Barril, X. (2014) Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites, *Journal of medicinal chemistry 57*, 8530-8539.

[100] (2016) PyMOL 1.8.4.0, Schrodinger.

[101] Jarlier, V., Nicolas, M. H., Fournier, G., and Philippon, A. (1988) Extended broad-spectrum beta-lactamases conferring transferable resistance to newer beta-lactam agents in Enterobacteriaceae: hospital prevalence and susceptibility patterns, *Reviews of infectious diseases 10*, 867-878.

[102] Livermore, D. M., Canton, R., Gniadkowski, M., Nordmann, P., Rossolini, G. M., Arlet, G., Ayala, J., Coque, T. M., Kern-Zdanowicz, I., Luzzaro, F., Poirel, L., and Woodford, N. (2007) CTX-M: changing the face of ESBLs in Europe, *Journal of Antimicrobial Chemotherapy 59*, 165-174.

[103] Hooton, T. M. (2012) Uncomplicated Urinary Tract Infection, *New England Journal of Medicine 366*, 1028-1037.

[104] Johnson, J. R., Johnston, B., Clabots, C., Kuskowski, M. A., and Castanheira, M. (2010) Escherichia coli sequence type ST131 as the major cause of serious multidrug-resistant E. coli infections in the United States, *Clinical infectious diseases 51*, 286-294.

[105] Kallen, A. J., Welch, H., and Sirovich, B. E. (2006) Current antibiotic therapy for isolated urinary tract infections in women, *Archives of internal medicine 166*, 635-639.

[106] Rogers, B. A., Sidjabat, H. E., and Paterson, D. L. (2011) Escherichia coli O25b-ST131: a pandemic, multiresistant, community-associated strain, *Journal of Antimicrobial Chemotherapy 66*, 1-14.

[107] Coque, T. M., Novais, Â., Carattoli, A., Poirel, L., Pitout, J., Peixe, L., Baquero, F., Cantón, R., and Nordmann, P. (2008) Dissemination of Clonally Related Escherichia coli Strains Expressing Extended-Spectrum β-Lactamase CTX-M-15, *Emerging Infectious Diseases 14*, 195-200.

[108] Kang, C. I., Cha, M. K., Kim, S. H., Ko, K. S., Wi, Y. M., Chung, D. R., Peck, K. R., Lee, N. Y., and Song, J. H. (2013) Clinical and molecular epidemiology of community-onset bacteremia caused by extended-spectrum beta-lactamase-producing Escherichia coli over a 6-year period, *Journal of Korean medical science 28*, 998-1004.

[109] Peirano, G., and Pitout, J. D. D. (2010) Molecular epidemiology of Escherichia coli producing CTX-M β-lactamases: the worldwide emergence of clone ST131 O25:H4, *International Journal of Antimicrobial Agents 35*, 316-321.

[110] Nicolas-Chanoine, M. H., Bertrand, X., and Madec, J. Y. (2014) Escherichia coli ST131, an intriguing clonal group, *Clinical microbiology reviews 27*, 543-574.

[111] Clermont, O., Dhanji, H., Upton, M., Gibreel, T., Fox, A., Boyd, D., Mulvey, M. R., Nordmann, P., Ruppe, E., Sarthou, J. L., Frank, T., Vimont, S., Arlet, G., Branger, C., Woodford, N., and Denamur, E. (2009) Rapid detection of the O25b-ST131 clone of Escherichia coli encompassing the CTX-M-15-producing strains, *Journal of Antimicrobial Chemotherapy 64*, 274-277.

[112] Weissman, S. J., Johnson, J. R., Tchesnokova, V., Billig, M., Dykhuizen, D., Riddell, K., Rogers, P., Qin, X., Butler-Wu, S., Cookson, B. T., Fang, F. C., Scholes, D., Chattopadhyay, S., and Sokurenko, E. (2012) High-Resolution Two-Locus Clonal Typing of Extraintestinal Pathogenic Escherichia coli, *Applied and Environmental Microbiology 78*, 1353-1360.

[113] Clermont, O., Gordon, D., and Denamur, E. (2015) Guide to the various phylogenetic classification schemes for Escherichia coli and the correspondence among schemes, *Microbiology 161*, 980-988.

[114] Park, Y. S., Bae, I. K., Kim, J., Jeong, S. H., Hwang, S. S., Seo, Y. H., Cho, Y. K., Lee, K., and Kim, J. M. (2014) Risk factors and molecular epidemiology of community-onset extended-spectrum beta-lactamase-producing Escherichia coli bacteremia, *Yonsei medical journal 55*, 467-475.

[115] Pagani, L., Dell'Amico, E., Migliavacca, R., D'Andrea, M. M., Giacobone, E., Amicosante, G., Romero, E., and Rossolini, G. M. (2003) Multiple CTX-M-Type Extended-Spectrum β-Lactamases in Nosocomial Isolates of Enterobacteriaceae from a Hospital in Northern Italy, *Journal of Clinical Microbiology 41*, 4264-4269.

[116] Bush, K. P., Timothy. Jacoby, George. (2015) CTX-M-type B-lactamases, Lahey Clinic.

[117] Tchesnokova, V., Billig, M., Chattopadhyay, S., Linardopoulou, E., Aprikian, P., Roberts, P. L., Skrivankova, V., Johnston, B., Gileva, A., Igusheva, I., Toland, A., Riddell, K., Rogers, P., Qin, X., Butler-Wu, S., Cookson, B. T., Fang, F. C., Kahl, B., Price, L. B., Weissman, S. J., Limaye, A., Scholes, D., Johnson, J. R., and Sokurenko, E. V. (2013) Predictive Diagnostics for Escherichia coli Infections Based on the Clonal Association of Antimicrobial Resistance and Clinical Outcome, *Journal of Clinical Microbiology 51*, 2991-2999.

[118] Ko, Y. J., Moon, H. W., Hur, M., Park, C. M., Cho, S. E., and Yun, Y. M. (2013) Fecal carriage of extended-spectrum beta-lactamase-producing Enterobacteriaceae in Korean community and hospital settings, *Infection 41*, 9-13.

[119] Kim, B., Kim, J., Seo, M. R., Wie, S. H., Cho, Y. K., Lim, S. K., Lee, J. S., Kwon, K. T., Lee, H., Cheong, H. J., Park, D. W., Ryu, S. Y., Chung, M. H., Ki, M., and Pai, H. (2013) Clinical characteristics of community-acquired acute pyelonephritis caused by ESBL-producing pathogens in South Korea, *Infection 41*, 603-612.

[120] Park, S., Byun, J.-H., Choi, S.-M., Lee, D.-G., Kim, S.-H., Kwon, J.-C., Park, C., Choi, J.-H., and Yoo, J.-H. (2012) Molecular epidemiology of extended-spectrum beta-lactamase-producing Escherichia coli in the community and hospital in Korea: emergence of ST131 producing CTX-M-15, *BMC Infectious Diseases 12*, 1-11.

[121] Shin, J., Kim, D. H., and Ko, K. S. (2011) Comparison of CTX-M-14- and CTX-M-15-producing Escherichia coli and Klebsiella pneumoniae isolates from patients with bacteremia, *Journal of Infection 63*, 39-47.

[122] Paul, S., Linardopoulou, E. V., Billig, M., Tchesnokova, V., Price, L. B., Johnson, J. R., Chattopadhyay, S., and Sokurenko, E. V. (2013) Role of Homologous Recombination in Adaptive Diversification of Extraintestinal Escherichia coli, *Journal of Bacteriology 195*, 231-242.

[123] Brisse, S., Diancourt, L., Laouenan, C., Vigan, M., Caro, V., Arlet, G., Drieux, L., Leflon-Guibout, V., Mentre, F., Jarlier, V., and Nicolas-Chanoine, M. H. (2012) Phylogenetic distribution of CTX-M- and non-extended-spectrum-beta-lactamase-producing Escherichia coli isolates: group B2 isolates, except clone ST131, rarely produce CTX-M enzymes, *Journal of Clinical Microbiology 50*, 2974-2981.

[124] Lopez-Cerero, L., Navarro, M. D., Bellido, M., Martin-Pena, A., Vinas, L., Cisneros, J. M., Gomez-Langley, S. L., Sanchez-Monteseirin, H., Morales, I., Pascual, A., and Rodriguez-Bano, J. (2014) Escherichia coli belonging to the worldwide emerging epidemic clonal group O25b/ST131: risk factors and clinical implications, *Journal of Antimicrobial Chemotherapy 69*, 809-814.

[125] Price, L. B., Johnson, J. R., Aziz, M., Clabots, C., Johnston, B., Tchesnokova, V., Nordstrom, L., Billig, M., Chattopadhyay, S., Stegger, M., Andersen, P. S., Pearson, T., Riddell, K., Rogers, P., Scholes, D., Kahl, B., Keim, P., and Sokurenko, E. V. (2013) The Epidemic of Extended-Spectrum-β-Lactamase-Producing Escherichia coli ST131 Is Driven by a Single Highly Pathogenic Subclone, H30-Rx, *mBio 4*, 1-10.

[126] Kim, S.-Y., Park, Y.-J., Johnson, J. R., Yu, J. K., Kim, Y.-K., and Kim, Y. S. Prevalence and characteristics of Escherichia coli sequence type 131 and its H30 and H30Rx subclones: a multicenter study from Korea, *Diagnostic Microbiology and Infectious Disease* 84, 97-101.

[127] Petty, N. K., Ben Zakour, N. L., Stanton-Cook, M., Skippington, E., Totsika, M., Forde, B. M., Phan, M. D., Gomes Moriel, D., Peters, K. M., Davies, M., Rogers, B. A., Dougan, G., Rodriguez-Bano, J., Pascual, A., Pitout, J. D., Upton, M., Paterson, D. L., Walsh, T. R., Schembri, M. A., and Beatson, S. A. (2014) Global dissemination of a multidrug resistant Escherichia coli clone, *Proceedings of the National Academy of Sciences 111*, 5694-5699.

[128] Greer, E. L., and Shi, Y. (2012) Histone methylation: a dynamic mark in health, disease and inheritance, *Nature Reviews Genetics 13*, 343-357.

[129] Dillon, S. C., Zhang, X., Trievel, R. C., and Cheng, X. (2005) The SET-domain protein superfamily: protein lysine methyltransferases, *Genome Biology 6*, 227-227.

[130] Li, Y., Trojer, P., Xu, C.-F., Cheung, P., Kuo, A., Drury, W. J., Qiao, Q., Neubert, T. A., Xu, R.-M., Gozani, O., and Reinberg, D. (2009) The Target of the NSD Family of Histone Lysine Methyltransferases Depends on the Nature of the Substrate, *The Journal of biological chemistry 284*, 34283-34295.

[131] Wang, G. G., Cai, L., Pasillas, M. P., and Kamps, M. P. (2007) NUP98-NSD1 links H3K36 methylation to Hox-A gene activation and leukaemogenesis, *Nature Cell Biology 9*, 804-812.

[132] Angrand, P.-O., Apiou, F., Stewart, A. F., Dutrillaux, B., Losson, R., and Chambon, P. (2001) NSD3, a New SET Domain-Containing Gene, Maps to 8p12 and Is Amplified in Human Breast Cancer Cell Lines, *Genomics 74*, 79-88.

[133] Berdasco, M., Ropero, S., Setien, F., Fraga, M. F., Lapunzina, P., Losson, R., Alaminos, M., Cheung, N.-K., Rahman, N., and Esteller, M. (2009) Epigenetic inactivation of the Sotos overgrowth syndrome gene histone methyltransferase NSD1 in human neuroblastoma and glioma, *Proceedings of the National Academy of Sciences 106*, 21830-21835.

[134] Keats, J.J., Maxwell, C. A., Taylor, B. J., Hendzel, M. J., Chesi, M., Bergsagel, P. L., Larratt, L. M., Mant, M. J., Reiman, T., Belch, A. R., and Pilarski, L. M. (2005) Overexpression of transcripts originating from the MMSET locus characterizes all t(4;14)(p16;q32)-positive multiple myeloma patients, *Blood 105*, 4060-4069.

[135] Huggins, D. J., Sherman, W., and Tidor, B. (2012) Rational Approaches to Improving Selectivity in Drug Design, *Journal of medicinal chemistry 55*, 1424-1444.

[136] Zhang, X., Yang, Z., Khan, S. I., Horton, J. R., Tamaru, H., Selker, E. U., and Cheng, X. (2003) Structural basis for the product specificity of histone lysine methyltransferases, *Molecular Cell 12*, 177-185.

[137] Qiao, Q., Li, Y., Chen, Z., Wang, M., Reinberg, D., and Xu, R.-M. (2011) The structure of NSD1 reveals an autoregulatory mechanism underlying histone H3K36 methylation, *The Journal of biological chemistry 286*, 8361-8368.

[138] An, S., Yeo, K., Jeon, Y., and Song, J.-J. (2011) Crystal structure of the human histone methyltransferase ASH1L catalytic domain and its implications for the regulatory mechanism, *The Journal of biological chemistry 286*, 8369-8374.

[139] Zheng, W., Ibanez, G., Wu, H., Blum, G., Zeng, H., Dong, A., Li, F., Hajian, T., Allali-Hassani, A., Amaya, M. F., Siarheyeva, A., Yu, W., Brown, P. J., Schapira, M., Vedadi, M., Min, J., and Luo, M. (2012) Sinefungin derivatives as inhibitors and structure probes of protein lysine methyltransferase SETD2, *Journal of the American Chemical Society 134*, 18004-18014.

[140] Southall, S. M., Wong, P. S., Odho, Z., Roe, S. M., and Wilson, J. R. (2009) Structural basis for the requirement of additional factors for MLL1 SET domain activity and recognition of epigenetic marks, *Molecular Cell 33*, 181-191.

[141] Wu, H., Min, J., Lunin, V. V., Antoshenko, T., Dombrovski, L., Zeng, H., Allali-Hassani, A., Campagna-Slater, V., Vedadi, M., Arrowsmith, C. H., Plotnikov, A. N., and Schapira, M. (2010) Structural biology of human H3K9 methyltransferases, *PloS one 5*, e8570.

[142] Rogawski, D. S., Ndoj, J., Cho, H. J., Maillard, I., Grembecka, J., and Cierpicki, T. (2015) Two Loops Undergoing Concerted Dynamics Regulate the Activity of the ASH1L Histone Methyltransferase, *Biochemistry 54*, 5401-5413.

[143] Morishita, M., and di Luccio, E. (2011) Structural insights into the regulation and the recognition of histone marks by the SET domain of NSD1, *Biochemical and biophysical research communications 412*, 214-219.

[144] Morishita, M., Mevius, D., and di Luccio, E. (2014) In vitro histone lysine methylation by NSD1, NSD2/MMSET/WHSC1 and NSD3/WHSC1L, *BMC structural biology 14*, 1-13.

[145] Leach, A. R. (1996) *Molecular modelling: principles and applications*, Longman, Harlow, England.

[146] Chang, Y., Sun, L., Kokura, K., Horton, J. R., Fukuda, M., Espejo, A., Izumi, V., Koomen, J. M., Bedford, M. T., Zhang, X., Shinkai, Y., Fang, J., and Cheng, X. (2011) MPP8 mediates the interactions between DNA methyltransferase Dnmt3a and H3K9 methyltransferase GLP/G9a, *Nature communications 2*, 1-10.

[147] Allali-Hassani, A., Kuznetsova, E., Hajian, T., Wu, H., Dombrovski, L., Li, Y., Gräslund, S., Arrowsmith, C. H., Schapira, M., and Vedadi, M. (2014) A Basic Post-SET Extension of NSDs Is Essential for Nucleosome Binding In Vitro, *Journal of Biomolecular Screening 19*, 928-935.

[148] Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot, *Acta Crystallographica Section D 66*, 486-501.

[149] Case, D. A., Darden, T. A., Cheatham, T. E., III, Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Roberts, B., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K. F., Paesani, F., Vanicek, J., Liu, J., Wu, X., Brozell, S.

R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M.-J., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., and Kollman, P. A. (2010) AMBER 11, University of California, San Francisco.

[150] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004) Development and testing of a general amber force field, *Journal of Computational Chemistry 25*, 1157-1174.

[151] Jakalian, A., Jack, D. B., and Bayly, C. I. (2002) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation, *Journal of Computational Chemistry 23*, 1623-1641.

[152] Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011) PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions, *Journal of Chemical Theory and Computation 7*, 525-537.

[153] Søndergaard, C. R., Olsson, M. H. M., Rostkowski, M., and Jensen, J. H. (2011) Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values, *Journal of Chemical Theory and Computation 7*, 2284-2295.

[154] Zhang, X., and Bruice, T. C. (2007) Histone lysine methyltransferase SET7/9: formation of a water channel precedes each methyl transfer, *Biochemistry 46*, 14838-14844.

[155] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water, *The Journal of Chemical Physics 79*, 926-935.

[156] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. (1984) Molecular dynamics with coupling to an external bath, *The Journal of Chemical Physics 81*, 3684-3690.

[157] Roe, D. R., and Cheatham, T. E. (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data, *Journal of Chemical Theory and Computation 9*, 3084-3095.

[158] Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: Visual molecular dynamics, *Journal of Molecular Graphics 14*, 33-38.

[159] Rayasam, G. V., Wendling, O., Angrand, P. O., Mark, M., Niederreither, K., Song, L., Lerouge, T., Hager, G. L., Chambon, P., and Losson, R. (2003) NSD1 is essential for early post-implantation development and has a catalytically active SET domain, *The EMBO journal 22*, 3153-3163.

[160] Kubicek, S., O'Sullivan, R. J., August, E. M., Hickey, E. R., Zhang, Q., Teodoro, Miguel L., Rea, S., Mechtler, K., Kowalski, J. A., Homon, C. A., Kelly, T. A., and Jenuwein, T. (2007) Reversal of H3K9me2 by a Small-Molecule Inhibitor for the G9a Histone Methyltransferase, *Molecular cell 25*, 473-481.

[161] Chang, Y., Ganesh, T., Horton, J. R., Spannhoff, A., Liu, J., Sun, A., Zhang, X., Bedford, M. T., Shinkai, Y., Snyder, J. P., and Cheng, X. (2010) Adding a lysine mimic in the design of potent inhibitors of histone lysine methyltransferases, *Journal of molecular biology 400*, 1-7.

[162] Liu, F., Chen, X., Allali-Hassani, A., Quinn, A. M., Wigle, T. J., Wasney, G. A., Dong, A., Senisterra, G., Chau, I., Siarheyeva, A., Norris, J. L., Kireev, D. B., Jadhav, A., Herold, J. M., Janzen, W. P., Arrowsmith, C. H., Frye, S. V., Brown, P. J., Simeonov, A., Vedadi, M., and Jin, J. (2010) Protein lysine methyltransferase G9a inhibitors: design, synthesis, and structure activity relationships of 2,4-diamino-7-aminoalkoxy-quinazolines, *Journal of medicinal chemistry 53*, 5844-5857.

[163] Devkota, K., Lohse, B., Liu, Q., Wang, M.-W., Stærk, D., Berthelsen, J., and Clausen, R. P. (2014) Analogues of the Natural Product Sinefungin as Inhibitors of EHMT1 and EHMT2, *ACS Medicinal Chemistry Letters 5*, 293-297.

[164] Nguyen, K. T., Li, F., Poda, G., Smil, D., Vedadi, M., and Schapira, M. (2013) Strategy to Target the Substrate Binding site of SET Domain Protein Methyltransferases, *Journal of chemical information and modeling 53*, 681-691.

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

[166] Levinson, N. M., and Boxer, S. G. (2014) A conserved water-mediated hydrogen bond network defines bosutinib's kinase selectivity, *Nature Chemical Biology 10*, 127-132.

[167] Ross, G., Morris, G., and Biggin, P. (2012) Rapid and accurate prediction and scoring of water molecules in protein binding sites, *PloS one 7*, e32036.

[168] Beuming, T., Farid, R., and Sherman, W. (2009) High-energy water sites determine peptide binding affinity and specificity of PDZ domains, *Protein science 18*, 1609-1619.

[169] Pearlstein, R., Hu, Q.-Y., Zhou, J., Yowe, D., Levell, J., Dale, B., Kaushik, V., Daniels, D., Hanrahan, S., Sherman, W., and Abel, R. (2010) New hypotheses about the structure-function of proprotein convertase subtilisin/kexin type 9: analysis of the epidermal growth factor-like repeat A docking site using WaterMap, *Proteins 78*, 2571-2586.

[170] Christopher, H., Thijs, B., and Woody, S. (2010) Hydration Site Thermodynamics Explain SARs for Triazolylpurines Analogues Binding to the A2A Receptor, *ACS Medicinal Chemistry Letters 1*, 160-164.

[171] Robinson, D., Sherman, W., and Farid, R. (2010) Understanding kinase selectivity through energetic analysis of binding site waters, *ChemMedChem 5*, 618-627.

[172] Wilson, D. K., Bohren, K. M., Gabbay, K. H., and Quiocho, F. A. (1992) An unlikely sugar substrate site in the 1.65 A structure of the human aldose reductase holoenzyme implicated in diabetic complications, *Science 257*, 81-84.

[173] Stec, B., Holtz, K. M., Wojciechowski, C. L., and Kantrowitz, E. R. (2005) Structure of the wild-type TEM-1 beta-lactamase at 1.55 A and the mutant enzyme Ser70Ala at 2.1 A suggest the mode of noncovalent catalysis for the mutant enzyme, *Acta Crystallographica D 61*, 1072-1079.

[174] Patel, S., Vuillard, L., Cleasby, A., Murray, C. W., and Yon, J. (2004) Apo and inhibitor complex structures of BACE (beta-secretase), *Journal of molecular biology 343*, 407-416.

[175] Filippakopoulos, P., Picaud, S., Mangos, M., Keates, T., Lambert, J. P., Barsyte-Lovejoy, D., Felletar, I., Volkmer, R., Muller, S., Pawson, T., Gingras, A. C., Arrowsmith, C. H., and Knapp, S. (2012) Histone recognition and large-scale structural analysis of the human bromodomain family, *Cell 149*, 214-231.

[176] Li, R., Sirawaraporn, R., Chitnumsub, P., Sirawaraporn, W., Wooden, J., Athappilly, F., Turley, S., and Hol, W. G. (2000) Three-dimensional structure of M. tuberculosis dihydrofolate reductase reveals opportunities for the design of novel tuberculosis drugs, *Journal of molecular biology 295*, 307-323.

[177] Prodromou, C., Roe, S. M., Piper, P. W., and Pearl, L. H. (1997) A molecular clamp in the crystal structure of the N-terminal domain of the yeast Hsp90 chaperone, *Nature structural biology 4*, 477-482.

[178] Li, Q., Qi, J., Wu, Y., Kiyota, H., Tanaka, K., Suhara, Y., Ohrui, H., Suzuki, Y., Vavricka, C. J., and Gao, G. F. (2013) Functional and structural analysis of influenza virus neuraminidase N3 offers further insight into the mechanisms of oseltamivir resistance, *Journal of virology 87*, 10016-10024.

[179] Kishida, H., Unzai, S., Roper, D. I., Lloyd, A., Park, S. Y., and Tame, J. R. (2006) Crystal structure of penicillin binding protein 4 (dacB) from Escherichia coli, both in the native form and covalently linked to various antibiotics, *Biochemistry 45*, 783-792.

[180] James, M. N., and Sielecki, A. R. (1983) Structure and refinement of penicillopepsin at 1.8 A resolution, *Journal of molecular biology 163*, 299-361.

[181] Figueiredo, A. C., Clement, C. C., Zakia, S., Gingold, J., Philipp, M., and Pereira, P. J. (2012) Rational design and characterization of D-Phe-Pro-D-Arg-derived direct thrombin inhibitors, *PloS one 7*, e34354.

[182] Vincent, B. C., Arendall, W. B., Jeffrey, J. H., Daniel, A. K., Robert, M. I., Gary, J. K., Laura, W. M., Jane, S. R., and David, C. R. (2010) MolProbity: all-atom structure validation for macromolecular crystallography, *Acta crystallographica D 66, 12-21.*

[183] Ryde, U. (1995) Molecular dynamics simulations of alcohol dehydrogenase with a four- or five-coordinate catalytic zinc ion, *Proteins: Structure, Function, and Bioinformatics 21*, 40-56.

[184] Holmberg, N., Ryde, U., and Bulow, L. (1999) Redesign of the coenzyme specificity in L-lactate dehydrogenase from bacillus stearothermophilus using site-directed mutagenesis and media engineering, *Protein engineering 12*, 851-856.

[185] Case, D. A., Darden, T.A., Cheatham, T.E., Simmerling, C.L., Wang, J., Duke, R.E., Luo, R., Walker, R.C., Zhang, W., Merz, K.M., Roberts, B., Hayik, S., Roitberg, A., Seabra, G., Swails, J., Goetz, A.W., Kollossvary, I., Wong, K.F., Paesani, F., Vanicek, J., Wolf, R.M., Liu, J., Wu, X., Brozell, S.R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M.J., Cui, G., Roe, D.R., Mathews, D.H., Seetin, M.G., Salomon-Ferrer, R., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., and Kollman, P.A. (2012) AMBER 12, University of California, San Francisco.

[186] Andrea, T. A., Swope, W. C., and Andersen, H. C. (1983) The role of long ranged forces in determining the structure and properties of liquid water, *The Journal of Chemical Physics 79*, 4576-4584.

[187] Case, D. A., Babin, V., Berryman, J. T., Betz, R. M., Cai, Q., Cerutti, D. S., Cheatham, T. E. I., Darden, T. A., Duke, R. E., Gohlke, H., Goetz, A. W., Gusarov, S., Homeyer, N., Janowski, P., Kaus, J., Kolossváry, I., Kovalenko, A., Lee, T. S., LeGrand, S., Luchko, T., Luo, R., Madej, B., Merz, K. M., Paesani, F., Roe, D. R., Roitberg, A., Sagui, C., Salomon-Ferrer, R., Seabra, G., Simmerling, C. L., Smith, W., Swails, J., Walker, R. C., Wang, J., Wolf, R. M., Wu, X., and Kollman, P. A. (2014) AMBER 14, University of California, San Francisco.

[188] Damm, K. L., and Carlson, H. A. (2006) Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures, *Biophysical Journal 90*, 4558-4573.

[189] Sanschagrin, P. C., and Kuhn, L. A. (1998) Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity, *Protein Science 7*, 2054-2064.

[190] Zhang, L., Zhang, H., Zhao, Y., Li, Z., Chen, S., Zhai, J., Chen, Y., Xie, W., Wang, Z., Li, Q., Zheng, X., and Hu, X. (2013) Inhibitor selectivity between aldo-keto reductase superfamily members AKR1B10 and AKR1B1: role of Trp112 (Trp111), *FEBS Letters 587*, 3681-3686.

[191] Harrison, D. H., Bohren, K. M., Ringe, D., Petsko, G. A., and Gabbay, K. H. (1994) An anion binding site in human aldose reductase: mechanistic implications for the binding of citrate, cacodylate, and glucose 6-phosphate, *Biochemistry 33*, 2011-2020.

[192] El-Kabbani, O., Darmanin, C., Schneider, T. R., Hazemann, I., Ruiz, F., Oka, M., Joachimiak, A., Schulze-Briese, C., Tomizaki, T., Mitschler, A., and Podjarny, A. (2004) Ultrahigh resolution drug design. II. Atomic resolution structures of human aldose reductase holoenzyme complexed with Fidarestat and Minalrestat: implications for the binding of cyclic imide inhibitors, *Proteins 55*, 805-813.

[193] Steuber, H., Zentgraf, M., Podjarny, A., Heine, A., and Klebe, G. (2006) High-resolution crystal structure of aldose reductase complexed with the novel sulfonyl-pyridazinone inhibitor exhibiting an alternative active site anchoring group, *Journal of molecular biology 356*, 45-56.

[194] Steuber, H. (2011) An old NSAID revisited: crystal structure of aldose reductase in complex with sulindac at 1.0 A supports a novel mechanism for its anticancer and antiproliferative effects, *ChemMedChem 6*, 2155-2157.

[195] Steuber, H., Zentgraf, M., La Motta, C., Sartini, S., Heine, A., and Klebe, G. (2007) Evidence for a novel binding site conformer of aldose reductase in ligand-bound state, *Journal of molecular biology 369*, 186-197.

[196] Brownlee, J. M., Carlson, E., Milne, A. C., Pape, E., and Harrison, D. H. (2006) Structural and thermodynamic studies of simple aldose reductase-inhibitor complexes, *Bioorganic chemistry 34*, 424-444.

[197] Carugo, O., and Bordo, D. (1999) How many water molecules can be detected by protein crystallography?, *Acta Crystallographica Section D 55*, 479-483.

[198] Balendiran, G. K., Sawaya, M. R., Schwarz, F. P., Ponniah, G., Cuckovich, R., Verma, M., and Cascio, D. (2011) The role of Cys-298 in aldose reductase function, *Journal of Biological Chemistry 286*, 6336-6344.

[199] Dineen, T. A., Chen, K., Cheng, A. C., Derakhchan, K., Epstein, O., Esmay, J., Hickman, D., Kreiman, C. E., Marx, I. E., Wahl, R. C., Wen, P. H., Weiss, M. M., Whittington, D. A., Wood, S., Fremeau, R. T., Jr., White, R. D., and Patel, V. F. (2014) Inhibitors of beta-Site Amyloid Precursor Protein Cleaving Enzyme (BACE1): Identification of (S)-7-(2-

Fluoropyridin-3-yl)-3-((3-methyloxetan-3-yl)ethynyl)-5'H-spiro[chromeno[2,3-b]pyridine-5,4'-oxazol]-2'-amine (AMG-8718), *Journal of medicinal chemistry 57*, 9811-9831.

[200] Barman, A., and Prabhakar, R. (2012) Protonation States of the Catalytic Dyad of β-Secretase (BACE1) in the Presence of Chemically Diverse Inhibitors: A Molecular Docking Study, *Journal of chemical information and modeling 52*, 1275-1287.

[201] Ghosh, A. K., Kumaragurubaran, N., Hong, L., Kulkarni, S. S., Xu, X., Chang, W., Weerasena, V., Turner, R., Koelsch, G., Bilcer, G., and Tang, J. (2007) Design, synthesis, and X-ray structure of potent memapsin 2 (beta-secretase) inhibitors with isophthalamide derivatives as the P2-P3-ligands, *Journal of medicinal chemistry 50*, 2399-2407.

[202] Brodney, M., Barreiro, G., Ogilvie, K., Hajos-Korcsok, E., Murray, J., Vajdos, F., Ambroise, C., Christoffersen, C., Fisher, K., Lanyon, L., Liu, J., Nolan, C., Withka, J., Borzilleri, K., Efremov, I., Oborski, C., Varghese, A., and O'Neill, B. (2012) Spirocyclic sulfamides as β-secretase 1 (BACE-1) inhibitors for the treatment of Alzheimer's disease: utilization of structure based drug design, WaterMap, and CNS penetration studies to identify centrally efficacious inhibitors, *Journal of medicinal chemistry 55*, 9224-9239.

[203] Bös, F., and Pleiss, J. (2008) Conserved Water Molecules Stabilize the Ω-Loop in Class A β-Lactamases, *Antimicrobial agents and chemotherapy 52*, 1072-1079.

[204] Maveyraud, L., Mourey, L., Kotra, L. P., Pedelacq, J.-D., Guillet, V., Mobashery, S., and Samama, J.-P. (1998) Structural Basis for Clinical Longevity of Carbapenem Antibiotics in the Face of Challenge by the Common Class A Beta-Lactamases from Antibiotic-Resistant Bacteria, *Journal of the American Chemical Society 120*, 9748-9752.

[205] Ness, S., Martin, R., Kindler, A. M., Paetzel, M., Gold, M., Jensen, S. E., Jones, J. B., and Strynadka, N. C. J. (2000) Structure-Based Design Guides the Improved Efficacy of Deacylation Transition State Analogue Inhibitors of TEM-1 β-Lactamase, *Biochemistry 39*, 5312-5321.

[206] Chung, C. W., Coste, H., White, J. H., Mirguet, O., Wilde, J., Gosmini, R. L., Delves, C., Magny, S. M., Woodward, R., Hughes, S. A., Boursier, E. V., Flynn, H., Bouillot, A. M., Bamborough, P., Brusq, J. M., Gellibert, F. J., Jones, E. J., Riou, A. M., Homes, P., Martin, S. L., Uings, I. J., Toum, J., Clement, C. A., Boullay, A. B., Grimley, R. L., Blandel, F. M., Prinjha, R. K., Lee, K., Kirilovsky, J., and Nicodeme, E. (2011) Discovery and characterization of small molecule inhibitors of the BET family bromodomains, *Journal of medicinal chemistry 54*, 3827-3838.

[207] Vollmuth, F., Blankenfeldt, W., and Geyer, M. (2009) Structures of the dual bromodomains of the P-TEFb-activating protein Brd4 at atomic resolution, *Journal of Biological Chemistry 284*, 36547-36556.

[208] Zhao, L. Cao, D., Chen, T., Wang, Y., Miao, Z., Xu, Y., Chen, W., Wang, X., Li, Y., Du, Z., Xiong, B., Li, J., Xu, C., Zhang, N., He, J., and Shen, J. (2013) Fragment-based drug discovery of 2-thiazolidinones as inhibitors of the histone reader BRD4 bromodomain, *Journal of medicinal chemistry 56*, 3833-3851.

[209] Ember, S. W., Zhu, J. Y., Olesen, S. H., Martin, M. P., Becker, A., Berndt, N., Georg, G. I., and Schonbrunn, E. (2014) Acetyl-lysine binding site of bromodomain-containing protein 4 (BRD4) interacts with diverse kinase inhibitors, *ACS Chemical Biology 9*, 1160-1171.

[210] Mastropaolo, D., Camerman, A., and Camerman, N. (1980) Folic Acid: Crystal Structure and Implications for Enzyme Binding, *Science 210*, 334-336.

[211] Trepel, J., Mollapour, M., Giaccone, G., and Neckers, L. (2010) Targeting the dynamic HSP90 complex in cancer, *Nature reviews cancer 10*, 537-549.

[212] Kung, P. P., Sinnema, P. J., Richardson, P., Hickey, M. J., Gajiwala, K. S., Wang, F., Huang, B., McClellan, G., Wang, J., Maegley, K., Bergqvist, S., Mehta, P. P., and Kania, R. (2011) Design strategies to target crystallographic waters applied to the Hsp90 molecular chaperone, *Bioorganic & medicinal chemistry letters 21*, 3557-3562.

[213] Millson, S. H., Chua, C. S., Roe, S. M., Polier, S., Solovieva, S., Pearl, L. H., Sim, T. S., Prodromou, C., and Piper, P. W. (2011) Features of the Streptomyces hygroscopicus HtpG reveal how partial geldanamycin resistance can arise with mutation to the ATP binding pocket of a eukaryotic Hsp90, *FASEB journal 25*, 3828-3837.

[214] Sharp, S. Y., Roe, S. M., Kazlauskas, E., Cikotiene, I., Workman, P., Matulis, D., and Prodromou, C. (2012) Co-crystalization and in vitro biological characterization of 5-aryl-4-(5-substituted-2-4-dihydroxyphenyl)-1,2,3-thiadiazole Hsp90 inhibitors, *PloS one 7*, e44642.

[215] Obermann, W. M., Sondermann, H., Russo, A. A., Pavletich, N. P., and Hartl, F. U. (1998) In vivo function of Hsp90 is dependent on ATP binding and ATP hydrolysis, *The Journal of cell biology 143*, 901-910.

[216] Samson, M., Pizzorno, A., Abed, Y., and Boivin, G. (2013) Influenza virus resistance to neuraminidase inhibitors, *Antiviralal Research 98*, 174-185.

[217] Vergara-Jaque, A., Poblete, H., Lee, E. H., Schulten, K., González-Nilo, F., and Chipot, C. (2012) Molecular Basis of Drug Resistance in A/H1N1 Virus, *Journal of chemical information and modeling 52*, 2650-2656.

[218] Massova, I., and Mobashery, S. (1998) Kinship and Diversification of Bacterial Penicillin-Binding Proteins and β-Lactamases, *Antimicrobial agents and chemotherapy 42*, 1-17.

[219] Prasad, B. V. L. S., and Suguna, K. (2002) Role of water molecules in the structure and function of aspartic proteinases, *Acta Crystallographica Section D 58*, 250-259.

[220] Khan, A. R., Parrish, J. C., Fraser, M. E., Smith, W. W., Bartlett, P. A., and James, M. N. G. (1998) Lowering the Entropic Barrier for Binding Conformationally Flexible Inhibitors to Enzymes, *Biochemistry 37*, 16839-16845.

[221] Wells, C. M., and Di Cera, E. (1992) Thrombin is a sodium ion activated enzyme, *Biochemistry 31*, 11721-11730.

[222] Di Cera, E., Guinto, E. R., Vindigni, A., Dang, Q. D., Ayala, Y. M., Wuyi, M., and Tulinsky, A. (1995) The Na+ Binding Site of Thrombin, *Journal of Biological Chemistry 270*, 22089-22092.

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

[224] Adrian, F. J., Ding, Q., Sim, T., Velentza, A., Sloan, C., Liu, Y., Zhang, G., Hur, W., Ding, S., Manley, P., Mestan, J., Fabbro, D., and Gray, N. S. (2006) Allosteric inhibitors of Bcr-abl-dependent cell proliferation, *Nature Chemical Biology 2*, 95-102.

[225] Xu, W., Doshi, A., Lei, M., Eck, M. J., and Harrison, S. C. (1999) Crystal structures of c-Src reveal features of its autoinhibitory mechanism, *Molecular Cell 3*, 629-638.

[226] Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S., and Walker, R. C. (2013) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald, *Journal of Chemical Theory and Computation 9*, 3878-3888.

[227] Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S., and Walker, R. C. (2012) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born, *Journal of Chemical Theory and Computation 8*, 1542-1555.

[228] Le Grand, S., Götz, A. W., and Walker, R. C. (2013) SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations, *Computer Physics Communications 184*, 374-380.

[229] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96),* 226-231.

[230] Schindler, T., Bornmann, W., Pellicena, P., Miller, W., Clarkson, B., and Kuriyan, J. (2000) Structural mechanism for STI-571 inhibition of abelson tyrosine kinase, *Science 289*, 1938-1942.

[231] Horio, T., Hamasaki, T., Inoue, T., Wakayama, T., Itou, S., Naito, H., Asaki, T., Hayase, H., and Niwa, T. (2007) Structural factors contributing to the Abl/Lyn dual inhibitory activity of 3-substituted benzamide derivatives, *Bioorganic & medicinal chemistry letters 17*, 2712-2717.

[232] Okram, B., Nagle, A., Adrian, F. J., Lee, C., Ren, P., Wang, X., Sim, T., Xie, Y., Wang, X., Xia, G., Spraggon, G., Warmuth, M., Liu, Y., and Gray, N. S. (2006) A general strategy for creating "inactive-conformation" abl inhibitors, *Chemistry & biology 13*, 779-786.

[233] Cowan-Jacob, S. W., Fendrich, G., Floersheimer, A., Furet, P., Liebetanz, J., Rummel, G., Rheinberger, P., Centeleghe, M., Fabbro, D., and Manley, P. W. (2007) Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia, *Acta Crystallographica D 63*, 80-93.

[234] Weisberg, E., Manley, P. W., Breitenstein, W., Bruggen, J., Cowan-Jacob, S. W., Ray, A., Huntly, B., Fabbro, D., Fendrich, G., Hall-Meyers, E., Kung, A. L., Mestan, J., Daley, G. Q., Callahan, L., Catley, L., Cavazza, C., Azam, M., Neuberg, D., Wright, R. D., Gilliland, D. G., and Griffin, J. D. (2005) Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl, *Cancer cell 7*, 129-141.

[235] Zhou, T., Commodore, L., Huang, W.-S., Wang, Y., Sawyer, T., Shakespeare, W., Clackson, T., Zhu, X., and Dalgarno, D. (2010) Structural analysis of DFG-in and DFG-out dual Src-Abl inhibitors sharing a common vinyl purine template, *Chemical biology & drug design 75*, 18-28.

[236] Jahnke, W., Grotzfeld, R. M., Pelle, X., Strauss, A., Fendrich, G., Cowan-Jacob, S. W., Cotesta, S., Fabbro, D., Furet, P., Mestan, J., and Marzinzik, A. L.. (2010) Binding or bending: distinction of allosteric Abl kinase agonists from antagonists by an NMR-based conformational assay, *Journal of the American Chemical Society 132*, 7043-7048.

[237] Zhou, T., Commodore, L., Huang, W. S., Wang, Y., Thomas, M., Keats, J., Xu, Q., Rivera, V. M., Shakespeare, W. C., Clackson, T., Dalgarno, D. C., and Zhu, X. (2011) Structural mechanism of the Pan-BCR-ABL inhibitor ponatinib (AP24534): lessons for overcoming kinase inhibitor resistance, *Chemical biology & drug design 77*, 1-11.

[238] Chan, W. W., Wise, S. C., Kaufman, M. D., Ahn, Y. M., Ensinger, C. L., Haack, T., Hood, M. M., Jones, J., Lord, J. W., Lu, W. P., Miller, D., Patt, W. C., Smith, B. D., Petillo, P. A., Rutkoski, T. J., Telikepalli, H., Vogeti, L., Yao, T., Chun, L., Clark, R., Evangelista, P., Gavrilescu, L. C., Lazarides, K., Zaleskas, V. M., Stewart, L. J., Van Etten, R. A., and Flynn, D. L. (2011) Conformational control inhibition of the BCR-ABL1 tyrosine kinase, including the gatekeeper T315I mutant, by the switch-control inhibitor DCC-2036, *Cancer cell 19*, 556-568.

[239] Weisberg, E., Choi, H. G., Ray, A., Barrett, R., Zhang, J., Sim, T., Zhou, W., Seeliger, M., Cameron, M., Azam, M., Fletcher, J. A., Debiec-Rychter, M., Mayeda, M., Moreno, D., Kung, A. L., Janne, P. A., Khosravi-Far, R., Melo, J. V., Manley, P. W., Adamia, S., Wu, C., Gray, N., and Griffin, J. D. (2010) Discovery of a small-molecule type II inhibitor of wild-type and gatekeeper mutants of BCR-ABL, PDGFRalpha, Kit, and Src kinases: novel type II inhibitor of gatekeeper mutants, *Blood 115*, 4206-4216.

[240] Liu, F., Wang, B., Wang, Q., Qi, Z., Chen, C., Kong, L. L., Chen, J. Y., Liu, X., Wang, A., Hu, C., Wang, W., Wang, H., Wu, F., Ruan, Y., Qi, S., Liu, J., Zou, F., Hu, Z., Wang, W., Wang, L., Zhang, S., Yun, C. H., Zhai, Z., Liu, J., and Liu, Q. (2016) Discovery and characterization of a novel potent type II native and mutant BCR-ABL inhibitor (CHMFL-074) for Chronic Myeloid Leukemia (CML), *Oncotarget* 7, 45562–45574.

[241] Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012) Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking, *Journal of medicinal chemistry 55*, 6582-6594.

[242] Teague, S. J., Davis, A. M., Leeson, P. D., and Oprea, T. (1999) The Design of Leadlike Combinatorial Libraries, *Angewandte Chemie 38*, 3743-3748.

[243] Meagher, K. L., Lerner, M. G., and Carlson, H. A. (2006) Refining the Multiple Protein Structure Pharmacophore Method: Consistency across Three Independent HIV-1 Protease Models, *Journal of medicinal chemistry 49*, 3478-3484.

[244] MacAuley, A., and Cooper, J. A. (1989) Structural differences between repressed and derepressed forms of p60c-src, *Molecular and Cellular Biology 9*, 2648-2656.

[245] Bakan, A., Nevins, N., Lakdawala, A., and Bahar, I. (2012) Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules, *Journal of Chemical Theory and Computation 8*, 2435-2447.

[246] Pereira de Jesus-Tran, K., Cote, P. L., Cantin, L., Blanchet, J., Labrie, F., and Breton, R. (2006) Comparison of crystal structures of human androgen receptor ligand-binding domain complexed with various agonists reveals molecular determinants responsible for binding affinity, *Protein Science 15*, 987-999.

[247] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters, *Proteins 65*, 712-725.

[248] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research 12*, 2825-2830.

[249] Comaniciu, D., and Meer, P. (2002) Mean Shift: A Robust Approach Toward Feature Space Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 603-619.

[250] Nagar, B., Hantschel, O., Young, M. A., Scheffzek, K., Veach, D., Bornmann, W., Clarkson, B., Superti-Furga, G., and Kuriyan, J. (2003) Structural basis for the autoinhibition of c-Abl tyrosine kinase, *Cell 112*, 859-871.

[251] Zhang, J., Adrian, F. J., Jahnke, W., Cowan-Jacob, S. W., Li, A. G., Iacob, R. E., Sim, T., Powers, J., Dierks, C., Sun, F., Guo, G. R., Ding, Q., Okram, B., Choi, Y., Wojciechowski, A., Deng, X., Liu, G., Fendrich, G., Strauss, A., Vajpai, N., Grzesiek, S., Tuntland, T., Liu, Y., Bursulaya, B., Azam, M., Manley, P. W., Engen, J. R., Daley, G. Q., Warmuth, M., and Gray, N. S. (2010) Targeting Bcr-Abl by combining allosteric with ATP-binding-site inhibitors, *Nature 463*, 501-506.

[252] Yang, J., Campobasso, N., Biju, M. P., Fisher, K., Pan, X. Q., Cottom, J., Galbraith, S., Ho, T., Zhang, H., Hong, X., Ward, P., Hofmann, G., Siegfried, B., Zappacosta, F., Washio, Y., Cao, P., Qu, J., Bertrand, S., Wang, D. Y., Head, M. S., Li, H., Moores, S., Lai, Z., Johanson, K., Burton, G., Erickson-Miller, C., Simpson, G., Tummino, P., Copeland, R. A., and Oliff, A. (2011) Discovery and characterization of a cell-permeable, small-molecule c-Abl kinase activator that binds to the myristoyl binding site, *Chemistry & biology 18*, 177-186.

[253] Gao, W., Bohl, C. E., and Dalton, J. T. (2005) Chemistry and Structural Biology of Androgen Receptor, *Chemical reviews 105*, 3352-3370.

[254] Estebanez-Perpina, E., Arnold, L. A., Nguyen, P., Rodrigues, E. D., Mar, E., Bateman, R., Pallai, P., Shokat, K. M., Baxter, J. D., Guy, R. K., Webb, P., and Fletterick, R. J. (2007) A surface on the androgen receptor that allosterically regulates coactivator binding, *Proceedings of the National Academy of Sciences 104*, 16074-16079.

[255] Nique, F., Hebbe, S., Peixoto, C., Annoot, D., Lefrancois, J. M., Duval, E., Michoux, L., Triballeau, N., Lemoullec, J. M., Mollat, P., Thauvin, M., Prange, T., Minet, D., Clement-Lacroix, P., Robin-Jagerschmidt, C., Fleury, D., Guedin, D., and Deprez, P. (2012) Discovery of diarylhydantoins as new selective androgen receptor modulators, *Journal of medicinal chemistry 55*, 8225-8235.

[256] Lin, X., Koelsch, G., Wu, S., Downs, D., Dashti, A., and Tang, J. (2000) Human aspartic protease memapsin 2 cleaves the β-secretase site of β-amyloid precursor protein, *Proceedings of the National Academy of Sciences 97*, 1456-1460.

[257] Hong, L., Koelsch, G., Lin, X., Wu, S., Terzyan, S., Ghosh, A. K., Zhang, X. C., and Tang, J. (2000) Structure of the Protease Domain of Memapsin 2 (β-Secretase) Complexed with Inhibitor, *Science 290*, 150-153.

[258] Hong, L., Turner, R. T., 3rd, Koelsch, G., Shin, D., Ghosh, A. K., and Tang, J. (2002) Crystal structure of memapsin 2 (beta-secretase) in complex with an inhibitor OM00-3, *Biochemistry 41*, 10963-10967.

[259] Turner, R. T., 3rd, Hong, L., Koelsch, G., Ghosh, A. K., and Tang, J. (2005) Structural locations and functional roles of new subsites S5, S6, and S7 in memapsin 2 (beta-secretase), *Biochemistry 44*, 105-112.

[260] May, P. C., Dean, R. A., Lowe, S. L., Martenyi, F., Sheehan, S. M., Boggs, L. N., Monk, S. A., Mathes, B. M., Mergott, D. J., Watson, B. M., Stout, S. L., Timm, D. E., Smith LaBell, E., Gonzales, C. R., Nakano, M., Jhee, S. S., Yen, M., Ereshefsky, L., Lindstrom, T. D., Calligaro, D. O., Cocke, P. J., Greg Hall, D., Friedrich, S., Citron, M., and Audia, J. E. (2011) Robust Central Reduction of Amyloid-β in Humans with an Orally Available, Non-Peptidic β-Secretase Inhibitor, *The Journal of Neuroscience 31*, 16507-16516.

[261] Schnell, J. R., Dyson, H. J., and Wright, P. E. (2004) Structure, dynamics, and catalytic function of dihydrofolate reductase, *Annual Review of Biophysics and Biomolecular Structure 33*, 119-140.

[262] Wu, Y. J., Guernon, J., Rajamani, R., Toyn, J. H., Ahlijanian, M. K., Albright, C. F., Muckelbauer, J., Chang, C., Camac, D., Macor, J. E., and Thompson, L. A. (2016) Discovery of furo[2,3-d][1,3]thiazinamines as beta amyloid cleaving enzyme-1 (BACE1) inhibitors, *Bioorganic & medicinal chemistry letters 26*, 5729-5731.

[263] Ghosh, A. K., Reddy, B. S., Yen, Y. C., Cardenas, E., Rao, K. V., Downs, D., Huang, X., Tang, J., and Mesecar, A. D. (2016) Design of Potent and Highly Selective Inhibitors for Human beta-Secretase 2 (Memapsin 1), a Target for Type 2 Diabetes, *Chemical science 7*, 3117-3122.

[264] Lam, T., Hilgers, M. T., Cunningham, M. L., Kwan, B. P., Nelson, K. J., Brown-Driver, V., Ong, V., Trzoss, M., Hough, G., Shaw, K. J., and Finn, J. (2014) Structure-based design of new dihydrofolate reductase antibacterial agents: 7-(benzimidazol-1-yl)-2,4-diaminoquinazolines, *Journal of medicinal chemistry 57*, 651-668.

[265] Graham, S. E., Zhang, L., Ali, I., Cho, Y. K., Ismail, M. D., Carlson, H. A., and Foxman, B. (2016) Prevalence of CTX-M extended-spectrum beta-lactamases and sequence type 131 in Korean blood, urine, and rectal Escherichia coli isolates, *Infection, genetics and evolution 41*, 292-295.

[266] Graham, S. E., Tweedy, S. E., and Carlson, H. A. (2016) Dynamic behavior of the post-SET loop region of NSD1: Implications for histone binding and drug development, *Protein Science 25*, 1021-1029.

[267] Martinez, L., Andrade, R., Birgin, E. G., and Martinez, J. M. (2009) PACKMOL: a package for building initial configurations for molecular dynamics simulations, *Journal of Computational Chemistry 30*, 2157-2164.

[268] Kalenkiewicz, A., Grant, B., and Yang, C.-Y. (2015) Enrichment of Druggable Conformations from Apo Protein Structures Using Cosolvent-Accelerated Molecular Dynamics, *Biology 4*, 344-366.

[269] Nagar, B., Bornmann, W. G., Pellicena, P., Schindler, T., Veach, D. R., Miller, W. T., Clarkson, B., and Kuriyan, J. (2002) Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571), *Cancer Research 62*, 4236-4243.

[270] Komander, D., and Rape, M. (2012) The ubiquitin code, *Annual review of biochemistry 81*, 203-229.

[271] Chau, V., Tobias, J. W., Bachmair, A., Marriott, D., Ecker, D. J., Gonda, D. K., and Varshavsky, A. (1989) A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein, *Science 243*, 1576-1583.

[272] Murtaza, M., Jolly, L. A., Gecz, J., and Wood, S. A. (2015) La FAM fatale: USP9X in development and disease, *Cellular and molecular life sciences 72*, 2075-2089.

[273] Schwickart, M., Huang, X., Lill, J. R., Liu, J., Ferrando, R., French, D. M., Maecker, H., O'Rourke, K., Bazan, F., Eastham-Anderson, J., Yue, P., Dornan, D., Huang, D. C. S., and Dixit, V. M. (2010) Deubiquitinase USP9X stabilizes MCL1 and promotes tumour cell survival, *Nature 463*, 103-107.

[274] Kapuria, V., Peterson, L. F., Fang, D., Bornmann, W. G., Talpaz, M., and Donato, N. J. (2010) Deubiquitinase Inhibition by Small-Molecule WP1130 Triggers Aggresome Formation and Tumor Cell Apoptosis, *Cancer Research 70*, 9265-9276.

[275] Pal, A., Young, M. A., and Donato, N. J. (2014) Emerging Potential of Therapeutic Targeting of Ubiquitin-Specific Proteases in the Treatment of Cancer, *Cancer Research 74*, 4955-4966.

[276] Komander, D., Clague, M. J., and Urbe, S. (2009) Breaking the chains: structure and function of the deubiquitinases, *Nature Reviews Molecular Cell Biology 10*, 550-563.

[277] Ritorto, M. S., Ewan, R., Perez-Oliva, A. B., Knebel, A., Buhrlage, S. J., Wightman, M., Kelly, S. M., Wood, N. T., Virdee, S., Gray, N. S., Morrice, N. A., Alessi, D. R., and Trost, M. (2014) Screening of DUB activity and specificity by MALDI-TOF mass spectrometry, *Nature communications 5*, 1-11.

[278] Al-Hakim, Abdallah K., Zagorska, A., Chapman, L., Deak, M., Peggie, M., and Alessi, Dario R. (2008) Control of AMPK-related kinases by USP9X and atypical Lys29/Lys33-linked polyubiquitin chains, *Biochemical Journal 411*, 249-260.

[279] Ernst, A., Avvakumov, G., Tong, J., Fan, Y., Zhao, Y., Alberts, P., Persaud, A., Walker, J. R., Neculai, A.-M., Neculai, D., Vorobyov, A., Garg, P., Beatty, L., Chan, P.-K., Juang, Y.-C., Landry, M.-C., Yeh, C., Zeqiraj, E., Karamboulas, K., Allali-Hassani, A., Vedadi, M., Tyers, M., Moffat, J., Sicheri, F., Pelletier, L., Durocher, D., Raught, B., Rotin, D., Yang, J., Moran, M. F., Dhe-Paganon, S., and Sidhu, S. S. (2013) A Strategy for Modulation of Enzymes in the Ubiquitin System, *Science 339*, 590-595.

[280] Dar, A. C., and Shokat, K. M. (2011) The evolution of protein kinase inhibitors from antagonists to agonists of cellular signaling, *Annual review of biochemistry 80*, 769-795.

[281] Meng, Y., Lin, Y.-l., and Roux, B. (2015) Computational Study of the "DFG-Flip" Conformational Transition in c-Abl and c-Src Tyrosine Kinases, *The Journal of Physical Chemistry B 119*, 1443-1456.

[282] Shan, Y., Seeliger, M. A., Eastwood, M. P., Frank, F., Xu, H., Jensen, M. O., Dror, R. O., Kuriyan, J., and Shaw, D. E. (2009) A conserved protonation-dependent switch controls drug binding in the Abl kinase, *Proceedings of the National Academy of Sciences 106*, 139-144.

[283] Filomia, F., De Rienzo, F., and Menziani, M. C. (2010) Insights into MAPK p38alpha DFG flip mechanism by accelerated molecular dynamics, *Bioorganic & medicinal chemistry 18*, 6805-6812.

[284] Vashisth, H., Maragliano, L., and Abrams, C. F. (2012) "DFG-flip" in the insulin receptor kinase is facilitated by a helical intermediate state of the activation loop, *Biophysical Journal 102*, 1979-1987.

[285] Oleinikovas, V., Saladino, G., Cossins, B. P., and Gervasio, F. L. (2016) Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations, *Journal of the American Chemical Society 138*, 14257-14263.