Exploring Binding Ensembles of Protein-Ligand Interactions in Explicit Solvent by Expanded Ensemble Simulation

A Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

in partial fulfillment

of the requirements for the degree

Master of Science

by

James Tyler Prillaman

May

2016

APPROVAL SHEET

The thesis

is submitted in partial fulfillment of the requirements

for the degree of

Master of Science

James Tyler Prillaman

AUTHOR

The thesis has been read and approved by the examining committee:

Dr. Micheal R. Shirts

Advisor

Dr. Geoffrey M. Geise

Dr. Kyle Lampe

Accepted for the School of Engineering and Applied Science:

James H. Ayl

Dean, School of Engineering and Applied Science

May

2016

Abstract

Binding interactions to proteins are important to consider in design of small molecules as putative drugs. We present a method for computationally generating and grouping sets of representative bounds structures of ligands and apply it to a model protein, T4-lysozyme L99A. Expanded ensemble simulations using explicitly solvated structures provides increased accuracy relative to continuum solvent representation while taking advantage of the enhanced sampling due to alchemical modification of the ligand. Enhanced sampling reduces the time needed for the ligand to experience many transition events and therefore sample most physically accessible locations around the protein. Alchemical modification further allows us to estimate binding free energies without significantly increasing simulation run times. We study four ligands with a range of known experimental binding affinities to identify alternative binding locations, even those of low ligand affinity. We present binding free energies to each location and to the protein overall using the multistate Bennett acceptance ratio method to estimate free energy differences between alchemical states. We find that contrary to expectations implicit solvent results were better able to estimate binding free energies for benzene both to the protein overall and to the experimental binding location. Explicit solvent simulations predict free energies that are not within error of the experimental measurements for benzene and phenol. Identified binding locations do not consistently match between implicit and explicit methods. Several issues identified during the simulations likely contributed to the disagreement, and we explain some ways that these issues can be addressed.

Acknowledgments

I would like to thank Dr. Michael Shirts for providing me with the resources and guidance needed to develop this work into a thesis. I also want to thank the members of the Shirts group for their patience in answering my questions on topics ranging from thermodynamics to computer operating systems. I appreciate the help of Levi Naden in catching typographical errors and improving the overall presentation of ideas in the written thesis. This work was completed with the assistance and cooperation of Kai Wang, an author on the paper establishing the precedent for the implicit solvent method of studying binding ensembles, whose code formed the basis for the analysis used to generate the results presented here. Special thanks to Drew Biedermann, who helped me to set up the HREMD simulations for explicit solvent study. I appreciate the flexibility and understanding of my Masters defense committee members Dr. Micheal R. Shirts (advisor), Dr. Kyle Lampe, and Dr. Geoffrey M. Geise (chairperson) in setting a defense date. The Jeffress Grant provided most of the funding for this work. The Rivanna computing cluster at the University of Virginia carried out all simulation and numerical analysis of the systems.

Contents

List of Figures

List of Tables

Introduction and Background

This study combines multiple computational methods to explore alternative binding locations of ligands to proteins. We start by explaining the concept of alternative binding locations, and how they can be helpful to modern drug development. We then review some of the key concepts that make the study practical, such as how alchemical modification 1) allows simulation to generate useful samples faster, and 2) provides a thermodynamic path by which we estimate binding free energies. The Methods portion (see Chapter 3) then explains some of the modifications we have made to collect meaningful results, as well as the complications we have identified. The Results portion (Chapter 4) then presents the data we collect for the test system T4-lysozyme with four model ligands.

1.1 Binding Locations

A binding location includes more than just one configuration of a ligand bound to a receptor molecule. Binding locations are regions of space within or next to a receptor to which the ligand binds. The overall free energy of the ligand-protein system is decreased when the ligand moves into these locations. Binding locations are often shown in literature with just a static image of a receptor and a molecule within the location. However, binding locations are more accurately represented by an ensemble of possible configurations that fully describe a molecule's ability to move around its mean position and reconfigure intramolecular degrees of freedom.

Specific and non-specific binding both affect the efficacy of a drug. Efficacy is the ability of a drug to perform its designed function. Specific binding is the series of interactions between a drug and its intended biological target, for example a site on a ion-channel membrane controlling the release of calcium. In this example, the drug may bind to an active site to release calcium, or it may bind to an allosteric site which then allows another molecule to bind and cause the release. The drug may also be an antagonist, where it binds to the receptor which prevents the binding of another molecule which would cause calcium to release. Non-specific binding is any association of the ligand to anything other than the specific binding site(s) [1]. This includes other areas of the same receptor molecule and locations on other molecules in the body.

Human serum albumin (HSA) is one example of a protein to which non-specific binding has a significant role in efficacy for several drugs, including warfarin [2] and diazepam [3]. Warfarin has a specific role as a vitamin-K antagonist anticoagulant, disrupting the natural cascade of protein formation that leads to blood coagulation [4]. Diazepam is a benzodiazapine, which is a class of drugs that are allosteric binders of GABAA receptors [5]. HSA typically binds fatty acids to deliver them throughout the body despite the aqueous environment [6]. HSA has multiple binding locations; Fasano et al. [7] show the seven different fatty acid binding locations within the protein, three of which are known to bind drug molecules. HSA acts as a non-specific target that helps transport these drugs through the body, but also keeps the drugs occupied limiting how much drug is available to bind. Drug design is most efficient when both binding to the specific target and to important non-specific proteins such as HSA is considered. Binding to non-specific transport proteins must be tight enough to provide the benefits of transport, but not too tight to prevent the drug reaching its specific target.

1.2 Alternative Binding Locations

Alternative binding locations are "alternative" to main binding locations and may have any binding affinity, though typically small. Main binding locations are the several locations that have the highest probability of being occupied by the ligand and therefore identified by experiment. While more sensitive experiments can identify more binding locations, we are interested in addressing all possible binding locations. The ligand may specifically bind to a region due to the ligand structure and how well that fits into the shape of the residues of the protein. It may also associate with the protein for a number of reasons, including favorable charged interactions and the hydrophobic effect. In either case, the arrangement of residues creates the environment in which the ligand and solvent interact. As long as there is a larger total increase in entropy and/or favorable enthalpic interactions for the ligand interacting with protein rather than with the solvent, binding will occur. The larger a protein, the higher the potential to have multiple binding locations. The collection of all possible configurations from all binding locations is the ligand binding ensemble.

Experimental methods are not always able to capture details of alternative binding locations. Palmer and Niwa [8] discuss several methods for determining ligand interactions with protein by x-ray crystallography. Crystallography requires growing a crystal of protein which is either already bound to ligand or is subsequently soaked in a ligand solution. The crystal is bombarded by x-rays, which scatter off of the electrons of the atoms in the crystal to create a two-dimensional diffraction pattern that provides high-resolution, though not complete, structural information about the arrangement of atoms in the protein-ligand complex. Subsequent refinement of the diffraction pattern using a model leads to assignment of structural attributes such as where the ligand molecules are located within the protein crystal lattice. Wielens et al. [9] explain some of the difficulties in performing crystallography, including inability to form a crystal under biological conditions. Further, we note that refinement typically assumes residues and the ligand molecule have only one or two locations in the structure. It is generally accepted that there is a threshold occupancy for ligands bound to the crystal, at approximately 20% occupancy. Above this threshold, the ligand crystal density is often possible to distinguish from the noise of the experiment, which is a consequence of the ligand having the same location within many of the proteins in the crystal. Below this threshold, however, the ligand is effectively invisible, and is a consequence of the ligand only binding to a certain fraction of proteins in the crystal. Additionally, there may be binding locations in solution that are not accessible in crystals because of the packing of proteins in the crystals. This implies that alternative binding locations may not be observed under a range of experimental conditions, especially for binding locations to which the ligand has low affinity. We refer to any location that can be well identified by crystallography as an experimental binding location.

Measurement of the free energy captures the effect of alternative binding locations, but can give misleading information if the energy is attributed to only a few locations. Isothermal titration calorimetry (ITC) is one method commonly used to determine the free energy of binding to a protein. Binding free energies express the affinity of a ligand to a receptor; more negative free energies indicate a more tightly bound ligand. Freire et al. [10] explain the fundamentals of isothermal titration calorimetry (ITC). Careful measurement of the heat released when either a solution of receptor or ligand is titrated by the other can provide both the enthalpic and entropic contributions to free energy. Mobley et al. [11] used both crystallography and ITC together to identify a binding location and its free energy. However, Freire et al. [10] note that the number of binding sites should be known when analyzing ITC data, because interpreting a signal that comes from multiple binding sites as from a single binding site will give incorrect results. Each site on the protein contributes to the free energy. The concept of alternative binding locations complicates the calculation as these sites are not necessarily independent: when a ligand is bound, this may prevent another ligand from binding to the same protein. ITC therefore provides an overall binding free energy, the free energy of binding to the protein over all sites together.

1.3 Virtual Drug Design

Detailed structural data extracted from orientations of a ligand in alternative binding locations enables better drug design. Anderson [12] outlines the general process of structure-based drug design (SBDD). SBDD is the process of building a drug that has a structure similar to molecules that bind the target receptor. A key step in this process is modifying a lead molecule to produce a binder with better properties. A similar approach is that of fragment-based lead discovery (FBLD), which is a motivating technique behind the experiments performed by Wielens et al. [9]. FBLD is the study of fragment molecules to identify which chemical moieties have affinity to which portions of a receptor molecule. These fragments are just small molecules, usually less than < 300 Da [13], and are typically portions of larger molecules known to bind to the receptor. The fragments can be combined to build drug lead molecules. Blundell et al. [14] discuss a crystallographic approach to high-throughput screening, and they mention that there is interest in studying low-affinity compounds as they may lead to the discovery of new binding modes. High-throughput screening (HTS) is the process of identifying the most viable candidates within a library of compounds. Both SBDD and FBLD build upon the idea of structure-activity relationships, that common structures will interact similarly with a receptor, and that the interaction is usually a direct result of the structural features such as functional groups. Bohacek and McMartin [15] discuss a framework for identification and display of important binding contacts once binding orientations are generated.

Experiment can only study drugs that have been created and are available in sufficient quantity. Researchers can instead study ligands that have yet to be synthesized using computational study. Ligands/fragments are typically based on existing drugs or binders which have been studied by experiment and for which the structures are well known. A computational study can explore the novel combinations of fragments to evaluate if each combination yields the desired binding activity. This means that synthetic methods only need to be developed for the most promising novel ligands, which saves time and material expense. Wang et al. [16] conclude that computation holds a key role within the experimental process, noting that virtual methods cannot replace scientific intuition, but also explaining how computation can reduce the risk associated with the creation of a new molecule.

This thesis develops and tests a computational approach that provides structural information of a ligand and receptor. Knowing which interactions are present and how strong are those interactions is useful to a variety of methods for drug design, from evaluating existing molecules to building new leads. We retain atom-level detail of the binding systems so that functional groups on the ligand could be associated with residues of the protein. Our method allows the ligand to explore the entire protein, in an attempt to find each alternative binding location, even those with low affinity for the ligand. We estimate the binding free energy to each of the binding locations as well as to the protein overall.

Two types of binding free energies of interest here are the absolute and relative free energies. An absolute binding free energy connects two physical ligand phases. It is the free energy difference of moving a ligand from the solvated (aqueous) phase to the bound (complexed) phase. Other reference phases could be used since free energy is a state function, notably the vacuum (low-density vapor) phase, but solvent is used as it is the most physically relevant reference. A relative binding free energy uses the absolute binding free energy of another ligand as a reference. Mathematically, a relative binding energy is equal to the difference in absolute binding free energies of two ligands, or

$$\Delta\Delta G_{target} = \Delta G_{target} - \Delta G_{ref} \tag{1.1}$$

where ΔG_{target} and ΔG_{ref} are the absolute binding free energies of the target and the reference ligand, respectively. Unless otherwise noted, all free energies reported in this study are absolute binding free energies.

Docking is a computational approach which matches the geometry and electrostatics of the ligand to the receptor, and is commonly used to quickly approximate which ligands in a set will bind to a receptor most tightly. Meng et al. [17] review the different methods currently in use to place the ligand within the protein, including fragment growth and Monte Carlo moves. The method will generate a set of ligand orientations relative to the receptor. These orientations are then ranked by a scoring function that estimates which structures have the most favorable binding interactions. Warren et al. [18] have noted that docking can identify orientations similar to crystallographic results.

Docking is not the best choice for the study of thermodynamically consistent ligand ensembles. Warren et al. [18] conclude that binding affinity predictions made by docking are not reliable in lead optimization. Lead optimization is the process of reducing a large set of interesting molecules down to a list of those most likely to have the desired binding affinity for the intended application. Totrov and Abagyan [19] explain how flexibility of the ligand and partial flexibility of the protein have been added to docking to improve performance, however, Blundell et al. [14] note that increased flexibility in docking can lead to worse predictions. Some scoring functions can consider entropy [17], but it is difficult without Monte Carlo (MC) or Molecular Dynamics (MD) methods to evaluate if orientations belong to a physical ensemble of bound structures. Generally, docking works best when binding activity is known *a priori*, and this information is incorporated into the docking algorithm [14, 18]. Docking still has a large speed advantage, able to analyze ligand sets in a matter of minutes [14], but its niche appears to be in generating initial structures for further study, as Mobley et al. [11] have done.

1.4 Molecular Dynamics and Binding Ensembles

Molecular dynamics captures the essential ligand motions to sample the correct ensemble of binding structures. MD simulates the physics of atomic motion by imposing a series of bonded and non-bonded interactions, the force-field, on each atom and moving them by integrating the equations of motion. The force-field parameters are chosen to so that the movement of atoms approximates the results of quantum mechanics. These simulations produce a series of configurations of the system, collectively called the trajectory. Each recorded configuration is a frame of the trajectory. During the molecular dynamics process, the ligand moves around the protein as it would in experimental solution, up to the accuracy of the force-field applied to the system. Frames from the trajectory are samples of the ligand's ensemble of configurations. Sampling, then, is the extent to which the samples of the trajectory represent the true ensemble. As long as the system is maintained at equilibrium, statistical mechanics can be used to describe the properties of the system.

Enhancements in computational speed, both by algorithm and by processing power, are welcome advancements towards virtual drug design. Ligand on- and off-rates for binding to a protein may be on time scales that require prohibitively long simulations to observe. Binding can occur on the order of milliseconds, while most simulations, even those performed on supercomputers, last less than a microsecond. We want to observe more on- and off-events during simulations. Each on-event is like a statistical sample of where the ligand will spend its time around the protein; many on-events must be observed to have confidence in how often each location is visited relative to the others. The next section explains how molecular dynamics can be used to gather more configurationally diverse members of a binding ensemble without increasing simulation run time.

1.5 Reducing Computational Effort with "Alchemistry"

Simulations that can change thermodynamic state as well as configuration can reduce the characteristic time scale for binding on- and off-events. These simulations periodically change state as a Monte Carlo move. Tempering is the process of moving between thermal equilibrium ensembles during a single simulation by changing the parameters under which the atoms interact. Lyubartsev et al. [20] present the equations for tempering, which ensure that the correct Boltzmann distribution is observed in each state. An increase in the temperature of the system encourages greater mean molecular speeds, which allows the simulation to capture rearrangement that would normally take far longer at the original temperature. The system can then cool down to sample at the desired temperature in this new arrangement. A change of state can also involve a non-thermal change in the Hamiltonian description of the system. Wang et al. [16] changed the Hamiltonian for a subset of the atoms in the system to increase the rates of ligand binding and unbinding in a study that was largely successful at predicting relative binding free energies. There are many modifications that can be made to a Hamiltonian. In this paper, we modify the non-bonded terms of the ligand molecule for the process of ligand binding.

Decreased interactions between the ligand and the environment increase how quickly the ligand can move during simulation. The protein residues and backbone present barriers to ligand movement, due to collision between molecules. Additional attraction due to favorable binding interactions also act to slow ligand diffusion by definition; the binding location is a site to which the ligand adsorbs. The binding locations are regions in which the ligand will have lower mean displacement due to the energetic barriers presented by the protein. The ligand can even become kinetically trapped, unable to leave during the span of a simulation. We observed 1-methylpyrrole oscillating for five nanoseconds within the buried binding location of lysozyme in our own preliminary simulation. We lower these energetic barriers directly by artificially scaling interaction terms within the force-field, specifically the Coulombic and Lennard-Jones terms. The ligand will fly without collision through the simulation space when these terms are scaled to zero.

Each set of interaction parameters defines a unique Hamiltonian. Interaction parameters, typically denoted as λ, control the level of interaction within simulation. We use interaction parameters to scale the Coulombic and Lennard-Jones terms, such that $\lambda = 0$ removes the interaction entirely. This modifies the ligand-environment interaction potential U_{inter} part of the Hamiltonian to be

$$\begin{aligned} U_{inter} &= U_c(r, \lambda_c) + U_{LJ}(r, \lambda_v) \\ &= \lambda_c \frac{q_1 q_2}{4\pi\epsilon_0\epsilon_r r} + 4\epsilon\lambda_v \left[\left(\frac{(1-\lambda_v)^2}{2} + \left(\frac{r}{\sigma}\right)^6 \right)^{-2} - \left(\frac{(1-\lambda_v)^2}{2} + \left(\frac{r}{\sigma}\right)^6 \right)^{-1} \right] \end{aligned} \tag{1.2}$$

where λ_c and λ_v are Coulombic and van der Waals parameters respectively. Variables q_1 and q_2 are atom charges, r is the distance between the two atoms, and $\epsilon_0\epsilon_r$ is the dielectric constant. The variables ϵ and σ are Lennard-Jones parameters, and the form of the Lennard-Jones equation is the soft-core potential presented by Beutler et al. [21] to avoid singularity at $r = 0$. A soft core model for Coulombic interaction does exist, but we avoid the need for it by decreasing the Coulombic parameter entirely before decreasing the van der Waals. Additionally, Naden and Shirts [22] show that choosing parameters in this order is statistically more efficient, providing lower free energy uncertainties; we explain the connection between interaction parameters and free energy estimates in Chapter 2. We therefore set $\lambda_c = 0$ for all $\lambda_v \neq 1$ for this study.

We define each unique nonphysical Hamiltonian as an “alchemical” state. Computational “alchemy” is a coined term that refers to changing the identity of atoms. An alchemical state then is any intermediate step in the transformation of one set of atoms into another. Atoms that feel lower interaction with their environment are alchemical; the atoms technically have a different identity even though they still have the same mass. Each level of interaction has its own ensemble of configurations available. The protein occupies a volume that is excluded from the ligand's ensemble when the interactions are left intact. The ligand can sample the entire simulation space with out any restrictions with interactions turned completely off. Each ensemble has a corresponding partition function, and by extension each level of interaction is a unique thermodynamic state. We specify an alchemical state by its interaction parameters.

Alchemical modification presents an additional variable with multiple degrees of freedom that can be used to increase the rate of sampling. The alchemical variable refers to the collection of states that are possible due to alchemical modification. The degrees of freedom related to this variable include the number of states and what parameters are assigned to each state. A simulation with five states that have evenly spaced interaction parameters may sample better than a simulation with ten states, eight of which have parameters close to 0. We can in principle optimize the number of states to tune simulation performance. In this study, we find that an increase in the number of states placed between existing states led to an increased number of binding events. It is not practical to optimize the choice of states for every ligand, but it is worth considering when sampling appears to be poor.

1.6 Select Simulation Methods for Enhanced Sampling

Two enhanced sampling methods are of interest to this study. An enhanced sampling method is any procedure that allows simulation to gather useful samples faster. A non-interacting ligand with atoms that have the same position as protein atoms is equally valid as any other configuration in the non-interacting state. That same configuration is very unlikely in the fully interacting state. These methods only change between states when the configuration is likely to belong to both ensembles.

Both methods take the same approach, but make changes in state by either a serial or parallel implementation. We use the name “expanded ensemble” for the serial implementation and Hamiltonian Replica Exchange Molecular Dynamics (HREMD) for the parallel.

We choose to use expanded ensemble as it has been shown to be more efficient than HREMD. Expanded ensemble simulation switches between states based only on the current energy of the simulation. Replica exchange refers to switching states between simulations, based on the energy difference between the two. Park [23] show that thermal tempering is more efficient when run as expanded ensemble rather than replica exchange. This is due to increased acceptance ratios in expanded ensemble. An acceptance ratio is the number of times a switch from one state to another was accepted over how many times it was rejected. Higher acceptance ratios generally indicate faster mixing in state space, that is, the temperature or Hamiltonian changes more rapidly. This carries over to mixing in configurational space. Quickly moving between states allows the ligand to become bound to the protein and subsequently leave on shorter time scales, since each state has its own off-rate. We refer to the ligand moving from the bound back to the mostly solvent areas of simulation space as a transition. This is much like a phase transition, where the phases are the bound and unbound configurational ensembles. Transitions are indirectly controlled by the rate of switching between states.

Hamiltonian Replica Exchange Molecular Dynamics uses a series of simulations that at any time are each in one of the predefined states. Wang et al. [24] used HREMD to perform their study of binding activity. The simulations are called replicas, as they each simulate a copy of the same system. Replicas run in parallel but have different interactions and positions at every step. The energies are compared between replicas on regular time intervals and the configurations are swapped between states so that each trajectory continues under a different level of interaction. These swaps are made with probability p_{ij} given by

$$p(i \leftrightarrow j) = p_{ij} = \min\left[1,\ \exp\left(\frac{\left(U_m(\vec{x}_j) - U_m(\vec{x}_i)\right) - \left(U_n(\vec{x}_j) - U_n(\vec{x}_i)\right)}{k_B T}\right)\right] \tag{1.3}$$

between replicas i and j in states m and n respectively. The equations for HREMD are reviewed by Okamoto [25] under the category multidimensional replica-exchange method. These swaps can be attempted multiple times before resuming simulation in order to promote better mixing of trajectories in state space.

Individual simulations in the expanded ensemble method do not depend on the energies of other simulations. We run multiple simulations at the same time to increase how quickly we generate samples, but in principle the same number of samples could be gathered from one very long expanded ensemble simulation. The concept of an expanded ensemble begins with the idea of a statistical ensemble. The canonical ensemble has the configuration integral Z_m,

$$Z_m = \frac{1}{N!}\int \exp\left(\frac{-U_m(\vec{x})}{k_B T}\right) \mathrm{d}\vec{x} \tag{1.4}$$

for state m with N number of particles. An expanded ensemble in alchemical simulation is the combination of each alchemical state's ensemble. The resulting expanded configuration integral Z is simply

$$Z = \sum_{m=1}^{k} Z_m \exp\left(W_m\right), \tag{1.5}$$

where W_m is a weighting factor. Lyubartsev et al. [20] give these formula for the case of temperature expanded ensemble. The decision to change state is a Monte

Carlo move, and at its simplest is evaluated as a jump to a neighboring state with probability given by the Metropolis condition such that

$$p(m \rightarrow n) = \min\left[1,\ \exp\left(\frac{\left(U_m(\vec{x_j}) - U_n(\vec{x_j})\right)}{k_B T}\right)\right]. \tag{1.6}$$

The factor $\exp(W_m)$ is chosen for each state to promote sampling in that state, since most systems will sample the states with the most favorable free energy according to the Boltzmann distribution and Metropolis condition. Metropolized-Gibbs is the algorithm used to perform the Monte Carlo moves in this study, and is explained in a later section (Sec. 3.1).

1.7 Solvent Representation

We explicitly include solvent molecules in our simulations as hydrophobic/hydrophilic effects require atom-level detail to fully capture binding interaction. Generalized Born (GB) is a model that represents the effect of solvent molecules as a spatially-dependent dielectric constant. GB and its application to calculating free energies within simulation by accessible surface area has been described by Still et al. [26], and later expanded to free energies using a volume integral by Labute [27]. The solvent is said to be implicit as the actual solvent molecules are not present during simulation. The implicit solvent models built using the GB equation are "statistical continuums" [26], such that the effect of solvent orientations are averaged. However, specific solvent orientation is speculated to be key in ligand-protein interaction. Bohacek and McMartin [15] suggests that non-covalent binding interactions can be described by hydrogen donor/acceptor ability for 90% of binding atoms. Zhang et al. [28] show an in-depth comparison of several implicit models with explicit solvation, including different solvent dielectrics. They specifically note that implicit solvent does not work well for certain types of protein regions, such as low-dielectric buried pockets. They attribute this inaccuracy in part to water-mediated hydrogen bonding. Anderson [12] gives a few ways in which solvent is an important part of ligand binding, including how displacement of a bound water molecule can increase the entropy of the system.

Treatment of solvent as individual molecules increases the computational demand of simulation. The program must loop over every atom to calculate interactions with the rest of the system. This expense is avoided when using continuum model for solvent, due to the fewer number of atoms. We hypothesize, however, that the increase in speed does not justify the reduced accuracy. We therefore evaluate the difference between binding results for previous implicit solvent simulation and a study of explicitly solvated systems.

We test the same system as Wang et al. [24] in order to understand how binding in implicit solvent is less accurate than explicit due to ignoring the ligand-solvent and protein-solvent interactions. The engineered protein T4-lysozyme mutant L99A has a large, buried, hydrophobic binding pocket, which is generally inaccessible by the solvent. The binding of several ligands to T4-lysozyme has been studied both experimentally and computationally [11], such that the binding free energies of the engineered pocket are well known. Wang et al. [24] were further able to identify multiple alternative binding locations and estimated the binding energy to each. Most of these alternative locations are on the surface of the protein, and likely interact with the solvent. Study of lysozyme interacting with each of the four ligands offers a unique test; the buried binding pocket represents a large kinetic barrier both to entrance and exit of the ligand. The ligands represent a range of binding affinities, from moderate to non-binding, and one ligand that requires flexibility of an internal residue in order to bind as observed in crystallographic experiment.

Theory

We quantify our results using the equations developed by statistical mechanics. We treat binding locations as volumes into which the ligand will partition according to the canonical ensemble. This allows us to estimate free energy differences based on the potential energies of samples gathered by simulation.

2.1 Ligand Binding Free Energies

Free energy is $k_B T$ times the logarithm of the number of microstates available to a system restricted by its environment. The canonical partition function includes both a configuration and a momenta integral. The configuration integral Z counts how many configurations are available, weighted by the energy of each configuration,

$$Z_V = \int\limits_V \exp\left(\frac{-U(\vec{x})}{k_B T}\right) \mathrm{d}\vec{x} \tag{2.1}$$

where k_B is the Boltzmann constant, T is the temperature, and $U(\vec{x})$ is the potential energy of a configuration within the phase space volume that is restricted by the physical volume V. Even though momentum is part of the phase space of the canonical ensemble, as long as atom masses do not change, we can assume that the contribution to the partition function is the same as in an ideal gas. It will therefore cancel out whenever we calculate differences in free energy where the number and identity of particles is conserved. Free energy G_V is directly related to the configuration integral,

$$G_V = -k_B T \ \ln Z_V, \tag{2.2}$$

where the ligand is restricted to volume V. If instead we define two volumes V_1 and V_2, we can explicitly define the free energy difference between the ligand occupying one volume relative to the other as the ratio of partition functions. The expression for the free energies and their difference are

$$\begin{aligned} G_{V_1} &= -k_B T \ \ln\left(Z_{V_1}\right) \\ G_{V_2} &= -k_B T \ \ln\left(Z_{V_2}\right) \\ \Delta G_{1-2} &= G_{V_2} - G_{V_1} \\ &= -k_B T \ \ln\left(\frac{Z_{V_2}}{Z_{V_1}}\right); \end{aligned} \tag{2.3}$$

this ratio is where the momentum contribution cancels. We also note that free energy differences can be defined not just between volumes but in general for any two states.

The alchemical states form a thermodynamic cycle. The (absolute) binding free energy is defined for the specific phase transition of a ligand from the solvated phase to the bound phase, where the ligand interacts predominantly with solvent and with protein respectively. We define a third, reference phase as a common point between two different systems. The solvent system consists of just ligand in solvent, while the protein system includes ligand, protein, and solvent. The ligand can be moved into the vacuum phase by removing all interactions with its environment in either system. Because the free energy is a state function, we can calculate the binding free energy indirectly by calculating the free energy of moving to the vacuum phase in one system and back from the vacuum phase in another.

The estimate for the binding free energy can be expressed simply,

$$\Delta G_{binding} = \Delta G_{coupling} + \Delta G_{volume} + \Delta G_{desolvation}, \tag{2.4}$$

where $\Delta G_{desolvation}$ is the free energy of removing the interactions of the ligand in the solvent system, $\Delta G_{coupling}$ is the free energy of introducing the interactions into the protein system, and ΔG_{volume} is the free energy of moving a vapor phase ligand from the solvent system volume to the protein system volume. This estimate of the binding free energy is defined in terms of the Gibbs free energy, at constant temperature and pressure.

The multistate Bennett acceptance ratio (MBAR) method produces free energy estimates of minimal uncertainty. Shirts and Chodera [29] derive MBAR using extended bridge sampling with an optimal estimator. For the case of uncorrelated samples taken from an ensemble obeying the Boltzmann distribution, the free energies can be written as an implicit function of themselves,

$$f_m = -\ln\left[\sum_{k=1}^{K}\sum_{n=1}^{N_k}\frac{\exp\left(-u_m(\vec{x}_{kn})\right)}{\sum_{j=1}^{K}\left[N_j\ \exp\left(f_j - u_j(\vec{x}_{kn}\right)\right]}\right]. \tag{2.5}$$

Here, $\vec{x}_{kn}$ is the n-th configuration from state k and N_k is the total number of configurations collected in state k. K is the total number of states, and f_m is the reduced free energy in state m,

$$f_m(\vec{x}) = F_m/k_B T_m, \tag{2.6}$$

with F_m being the appropriate free energy for the ensemble that is sampled. The

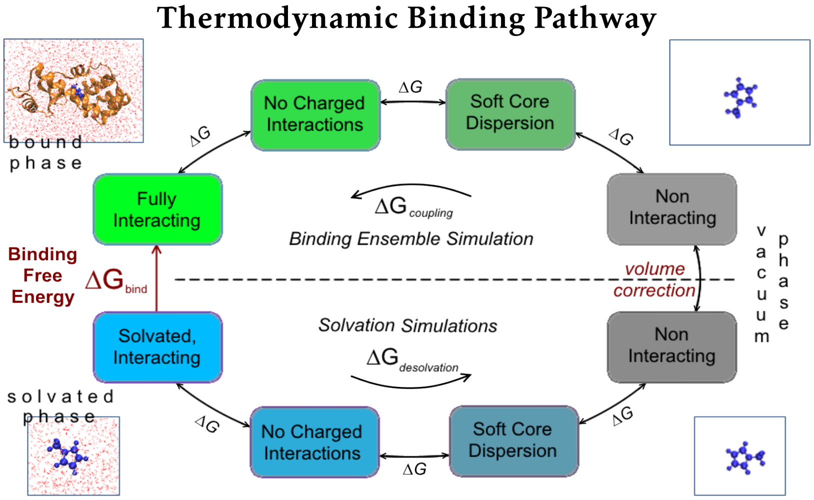


Figure 2.1: Alchemical modification provides a pathway through which we estimate binding free energies. Three phases are represented in the figure, with bound and solvated phases at left and the vacuum (or vapor) phase at the right; the binding free energy is the difference between the solvated and bound phases. The top and bottom halves of the cycle correspond to two different types of simulation; the simulations gather samples with the ligand in the protein system and the ligand in solvent respectively. Boxes represent alchemical states, and are labeled according to their level of interaction between the ligand and environment. The box colors fading to gray represent diminishing interaction of ligand with environment from left to right. The two inserts at top left and top right depict a ligand (in blue) interacting with the system and with no other molecules, respectively. The inserts at bottom depict only the ligand in solvent, with water represented by small red lines. The sum of the three terms $\Delta G_{desolvation}$, ΔG_{volume}, and $\Delta G_{coupling}$ is equal to the overall binding free energy, ΔG_{bind}.

reduced potential u_m is defined as

$$u_m(\vec{x}) = (U_m(\vec{x}) + P_m V(\vec{x}))/k_B T_m \tag{2.7}$$

for configurations in the canonical (NVT) and isobaric (NPT) ensembles which can be expanded by temperature (T_m), pressure (P_m), or Hamiltonian effecting just the potential energies ($U_m(\vec{x})$). $V(\vec{x})$ is the volume of configuration $\vec{x}$. The reduced potential can also be defined for grand-canonical ensembles (constant chemical potential μ). Solving Eqn. 2.5 by any method for systems of non-linear equations produces a set of free energy estimates. Further, MBAR can compare energies in states for which no configurations were generated, as long as the potential energy in that state was computed for all samples collected. We use the *pymbar* implementation of MBAR [30] to estimate the free energy differences between states of a simulation.

Simulation at constant volume is used to estimate free energies of constant pressure phase transitions. Due to numerical stability issues when sampling from multiple states, the protein-ligand simulations were carried out at constant volume (the NVT ensemble), instead of constant pressure (NPT). The solvent-ligand simulations are kept at a constant pressure, such that the box volume contracts when moving towards the vacuum phase; no correction is needed for $\Delta G_{solvation}$. This means that the free energies estimated for the NVT simulations are Helmholtz free energies. A correction from the Helmholtz (ΔA) to Gibbs (ΔG) free energies is needed, as well as defining the volume correction to standard state. These energies are related by $\Delta G = \Delta A + \Delta(PV)$. The difference is defined between the end states, that is, the fully coupled and the fully uncoupled lambda states. Each state will have a different pressure within the same volume, as the system loses a small amount of pressure due to the ligand becoming non-interacting. $\Delta G_{coupling}$ is technically defined between the end states at the same pressure, the pressure of the fully interacting system. We therefore estimate the Gibbs free energy of moving the ligand in the vacuum phase at the uncoupled pressure back to the pressure of the coupled state. The equation for the binding free energy including corrections is

$$\Delta G_{binding} = \left[\Delta A_{coupling} + (P_2 V_b - P_1 V_b) + \int_{P_2}^{P_1} \left(\frac{\mathrm{d}G}{\mathrm{d}P} \right)_T \mathrm{d}P \right] + k_B T \ln \left(\frac{V^\circ}{V_b} \right) + \Delta G_{desolvation}, \tag{2.8}$$

where P_1 and P_2 are the coupled and uncoupled state pressures respectively, $V^\circ = N_A^{-1}$ L is the standard volume, $V_b = 212$ nm^3 is the simulation volume, N_A is Avagadro's number, and $(\mathrm{d}G/\mathrm{d}P)|_T = V$ is applied assuming the system is a fluid with the properties of water. Gilson et al. [31] give a rigorous explanation of reference states and standard concentration, which leads to the form of the volume correction used here. We include Eqn. 2.8 as it represents the rigorous definition of the free energy difference. These simulations occur in water, which is effectively an incompressible fluid. This implies that large changes in pressure result in only minute changes in volume of the fluid. If we assume that the volume remains constant, the $\mathrm{d}(PV)$ and $\int V\mathrm{d}(P)$ terms cancel. These terms would be significant for a compressible fluid, but as the system is mostly water we will neglect the pressure corrections. The simplified version of the expression for the Gibbs free energy difference is

$$\Delta G_{binding} = \Delta A_{coupling} + k_B T \ln\left(\frac{V^\circ}{V_b}\right) + \Delta G_{desolvation}, \tag{2.9}$$

which is rigorous for a truly incompressible system.

We can define a binding free energy for each individual location. This is implied by the concept of non-specific binding, the idea that a ligand will have different affinity to different locations on the same receptor. We quantify the difference in affinity. Each location will have unique interactions with the ligand due to the residues and solvent molecules that define the boundaries of that location. The binding ensemble is formally the collection of all ligand configurations interacting with the protein, but we consider each location to have its own subensemble of configurations as part of the larger binding ensemble. These subensembles correspond to binding location configuration integrals,

$$Z_i = \int\limits_{V_i} \exp\left(\frac{-U_c(\vec{x})}{k_B T}\right) \mathrm{d}\vec{x}\,, \tag{2.10}$$

where V_i is the volume of the i-th binding location, and $\vec{x}$ was generated while in the fully-interacting state $m = c$.

We can identify binding locations by areas of high local density of coupled samples. This is an assumption based on the nature of the definition of free energy. As an example, take a free energy difference of -1.00 kcal/mol between two locations of the same volume. At 300 K, the ratio of partition functions is equal to 5.35, indicating a five times higher density of configurations in the lower free energy location. An increase in this probability density to 50, 100 and 1000 times more configurations at the same temperature corresponds to free energy differences of -2.33, -2.75, and -4.12 kcal/mol respectively.

We use spatial clustering to identify which samples from simulation belong to individual binding locations (specific details of clustering described in Sec. A.4). A three-dimensional grid is artificially imposed over these samples, so that each sample is located within one voxel of the grid. The volume of the binding location is the total volume of all voxels that have at least one sample from the cluster. Much of the interpretation of ensembles is originally presented by Wang et al. [24] and is expanded here for clarity. Free energies to each binding location are estimated to allow detailed comparison with implicit solvent results. We use MBAR to estimate these single location free energies, in addition to overall free energies. We combine all clustered samples from the fully interacting state with all samples from other states that fall into the cluster's voxels. The collection of samples are then subsampled by state and evaluated by MBAR to estimate the $\Delta G_{coupling}$ for that location. The single location binding free energies ΔG_i are then estimated by equation 2.9.

One measure of convergence of the simulations is to compare empirical binding location occupancies with the free energy estimates for those same locations. We define convergence of the simulations as the theoretical number of samples collected past which any additional samples will not change how many binding locations are identified or change the free energies by more than the estimated error of those energies. We define the occupancy as the ratio of the number of samples within a binding location to the total number in all binding locations. The occupancy O_i is given as

$$O_i = \frac{z_i}{\sum\limits_{n=1}^{N} z_n}, \tag{2.11}$$

where z_i is the number of samples within V_i, and there are N identified binding locations. The occupancy should be consistent with the ratio of partition functions determined from free energy estimates. An estimate of the occupancy O_i' can also be calculated from the free energy,

$$O_i' = \frac{Z_i}{\sum\limits_{n=1}^{N} Z_n} = \frac{\exp\left(-\Delta G_i/k_B T\right)}{\sum\limits_{n=1}^{N} \exp\left(-\Delta G_n/k_B T\right)}. \tag{2.12}$$

In the limit of infinite sampling, O_i should be equivalent to O_i'. A similar but alternative calculation is to estimate the free energy difference using the empirical occupancies. We choose the volume outside of the protein radius as a reference to define the free energy difference,

$$V_s = \frac{4}{3}\pi\left((r_p + 5\text{Å})^3 - r_p^3\right), \tag{2.13}$$

where r_p is the radius of the protein measured to the atom farthest from geometric center. Samples within this volume are predominantly interacting with solvent, and are a decent representation of the solvent phase. We use the number of samples z_i as an approximation to the partition function to estimate a free energy G_i' with arbitrary reference,

$$G_i' = -k_B T \ \ln(z_i). \tag{2.14}$$

The free energy difference between the solvent volume and the binding location volume is then

$$\begin{aligned}\Delta G_i' &= G_i' - G_s' - k_B T \ \ln \frac{V_s}{V^\circ} \\ &= -k_B T \ \ln z_i + k_B T \ \ln z_s - k_B T \ \ln \frac{V_s}{V^\circ} \\ &= -k_B T \ \ln \left(\frac{z_i}{z_s} \frac{V_s}{V^\circ} \right),\end{aligned} \tag{2.15}$$

adding in the free energy correction from the solvent volume to the standard reference volume. Similar to the occupancies, any difference between ΔG_i and $\Delta G_i'$ can indicate that sufficient sampling was not achieved.

Comparing free energies to already measured experimental values gives an indication of how well simulation is able to represent ligand binding. A simulation of any system with converged sampling, appropriate system variables (pressure, temperature, etc.), and accurate force-field should produce overall free energies that are within error of the experimental values. In both isothermal titration calorimetry and a converged simulation, the entire binding ensemble is sampled. We can also compare free energies for individual locations. Binding free energies of the experimentally observed binding location should agree with the free energy estimates determined by Mobley et al. [11]. In their experiment, the ligand was restrained to only sample the buried pocket, regardless of the alchemical state of the ligand. We expect these single location binding free energies to match for each ligand if we are able to identify the experimental location for that ligand.

Computational Methods

We have conducted simulations to study protein-ligand binding ensembles. This study uses explicit solvent expanded ensemble simulations to estimate binding activity, both binding locations and free energies. We compare results for previously studied systems of T4-lysozyme L99A, which have been treated experimentally and in both implicit and explicit solvent simulation. Minor optimization of parameters allowed us to collect data to present in the Results portion (see Chapter 4), but several unforeseen complications have led us to impose approximations in how the data is interpreted. Appendix A lists further details of how we perform these simulations.

3.1 System and Simulation Procedures

We simulate four ligands addressed by previous computational studies. Mobley et al. [11] show that the Val111 residue rotates only upon binding with *para*-xylene. Wang et al. [24] model this ligand to show how well flexibility is captured during simulation. They also model benzene, 1-methylpyrrole, and phenol as examples of a relatively strong binding ligand, a moderate ligand, and a non-binder. We mirror the analysis that Wang et al. [24] perform, in order to directly compare the results of implicit solvent to explicit solvent. Figure 3.1 shows these molecules along with

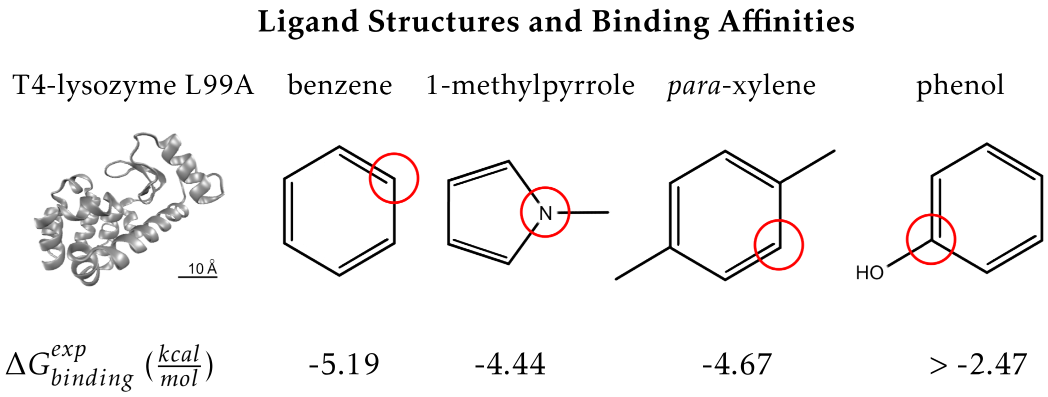

Ligand Structures and Binding Affinities
T4-lysozyme L99A
benzene
1-methylpyrrole
para-xylene
phenol
10 Å
N
HO
$\Delta G^{exp}_{binding}$ $(\frac{kcal}{mol})$
-5.19
-4.44
-4.67
> -2.47

Figure 3.1: We test the accuracy of expanded ensemble at identifying which ligands in a small set bind and to what extent. The Leu99Ala mutation in T4-lysozyme allows enough free volume for benzene-like molecules to bind [32]. The four ligands in the test set are shown with the atom calculated as closest to the center of geometry circled; this atom is used to track the ligand's trajectory. Note that for benzene and *para*-xylene, the choice of center atom is somewhat ambiguous due to their planes of symmetry. Experimental binding free energies are reported by Mobley et al. [11]. Rendered images in all figures are generated by VMD [33].

a cartoon representation of lysozyme.

Randomized initial configurations remove bias towards any particular location on the protein. We run preliminary simulations to generate initial configurations for each production simulation. These simulations run with the ligand always in the fully uncoupled state, so that the ligand can move to any location. We pull frames from these simulations to use as initial configurations. Starting from these random locations gives the set of simulations a better opportunity to find different binding locations with short run times. At long run times nearing and beyond convergence, the number of binding locations will remain the same, as they all will have been sampled by each simulation. Simulations that use these configurations start in the uncoupled state, so that the ligand does not immediately entangle other molecules and is able to travel for a stochastic amount of time before gathering physical samples.

Metropolized-Gibbs (MG) sampling in state space increases movement in configuration space. The protein-ligand simulations switch states by the MG algorithm. Chodera and Shirts [34] rigorously define the equations for MG as an example of independence sampling with rejection. The algorithm considers changing the state on regular intervals, typically shorter than the interval between collecting samples. We allow the state to change ten times between samples. Each state is assigned a probability by the equations

$$-u_n(\vec{x}_m) + w_n = \frac{-U_n(\vec{x}_m) + W_n}{k_B T}, \tag{3.1}$$

$$p_n = \frac{\exp\left(-u_n + w_n\right)}{\sum\limits_{k \neq m} \left[\exp\left(-u_k + w_k\right)\right]}, \tag{3.2}$$

where n is any state other than m, u_n is the reduced potential energy evaluated in state n of the current configuration from state m, and w_n is the reduced weight for state n. With these definitions $\sum\limits_{n \neq m} [p_n] = 1$, where $n \neq m$ implies summation over all K states in the range $[0,\ K-1]$ except state m. A new state is randomly suggested according to these probabilities, and then accepted with probability p_a in order to preserve detailed balance,

$$p_a = \min\left[1,\ \frac{\sum\limits_{k \neq m} \exp\left(-u_k + w_k\right)}{\sum\limits_{k \neq n} \exp\left(-u_k + w_k\right)}\right] \tag{3.3}$$

where $\min[1,\ \mathrm{y}]$ is the Metropolis function. Chodera and Shirts [34] explain how this algorithm encourages the state to change more frequently than independence sampling alone. Frequent changes in the state value allow the simulation to frequently sample in a more decoupled state between sampling in coupled states. This encourages movement of the ligand around the simulation between coupled samples.

The probabilities are weighted in order to increase frequency of sampling in each state. Lyubartsev et al. [20] suggests including these weights when describing the algorithm for temperature expanded ensemble. Without these weights, the simulations would quickly move to the most favorable state and remain there due to the exponential nature of the partition functions. With weights $\{W_0, W_1, ..., W_{K-1}\}$ equal to the free energy differences between states, the frequency of sampling in the limit of infinite sampling would be even across all states [34]. Equation 2.3 becomes

$$w_n = (G_n - G_0)/k_B T = \ln(Z_0/Z_n), \tag{3.4}$$

for the free energy of state n relative to state 0. The state selection probability in the MG algorithm (Eqn. 3.2) becomes

$$p_n = \frac{Z_0/Z_n \exp(-u_n)}{\sum\limits_{k \neq m} [Z_0/Z_k \exp(-u_k)]} = \frac{\exp(-u_n)}{\sum\limits_{k \neq m} [\exp(-u_k)\ Z_n/Z_k]}. \tag{3.5}$$

The denominator is shifted by the ratio of partition functions so that states with smaller Z_n have a larger probability of being chosen even when $\exp(-u_k)$ would be relatively large.

We employ the Wang-Landau algorithm [35] followed by iterative evaluation by MBAR to determine weights for the protein-ligand simulations. The Wang-Landau algorithm develops an estimate of the free energies by interactively updating the weights during a simulation (see Sec. A.3). We improve upon these weights by passing the energies for the latter half of a Wang-Landau simulation to MBAR. The algorithm works to obtain an even number of samples in each state, which in general decreases the uncertainty in MBAR evaluation relative to many samples from a single state with few in another. This follows from the similarity between MBAR and BAR [29] and the argument Bennett [36] presents related to spending equal time sampling each state, where samples are expected to take roughly the same amount of time independent of state. We use the first MBAR estimates to run a new simulation, which can then be evaluated by MBAR itself. This process should be repeated until the number of samples generated by a simulation is roughly equivalent. We perform the process for three evaluations of MBAR for benzene, *para*-xylene, and phenol, with 1-methylpyrrole having several more evaluations from preliminary simulation.

3.2 Alchemical Schedule and Weights

We want to analyze samples from the fully coupled state to determine binding locations, as the other states are increasingly non-physical. The ligand rapidly loses charge as the state index increases, having no charge by state 3. We were unable to collect samples in the most coupled state, state 0, due to a programming error in switching between states. Whenever the algorithm would suggest a switch into state 0, the value of the state was not properly updated, leaving the simulation in whichever state was previously visited. Instead of the physically relevant fully coupled samples, we collect state 1 samples from each of the twenty simulations and use these to identify the binding locations. This is an approximation, and creates a slightly undercharged molecule. This is expected to impact the most polar of the four ligands, phenol, to the greatest extent. We expect the occupancies to be effected the most by this change, as they will match the ensembles for state 1, where the overall and single location free energies are estimated to state 0. We examine the extent of the effects of this approximation in the results section.

Increasing the probability of sampling in the fully coupled state increases the number of physical samples. The weights W_m equal to free energy differences encourage even sampling as discussed in Sec. 3.1. The weights can be modified to increase the frequency of visiting a state relative to the others. We can increase the probability of sampling in a state by approximately a factor f_n. Adding $k_B T \ \ln(f_n)$ to W_n changes the selection probability as follows,

$$p_n = \frac{\exp\left((-u_n + w_n + \ln(f_n)\right)}{\sum\limits_{k \neq m} \left[\exp\left(-u_k + w_k + \ln(f_k)\right)\right]} = f_n \frac{\exp\left((-u_n + w_n\right)}{\sum\limits_{k \neq m} \left[f_k \exp\left(-u_k + w_k\right)\right]}. \tag{3.6}$$

Wang et al. [24] found for HREMD that multiple replicas in the coupled and uncoupled states are better for ligand sampling, since an increase in both states gathers more physical samples and presents more chances to move freely within a set amount of simulation time. This necessarily decreases the number of samples in all the states between the fully coupled and uncoupled states. The factors f_n allow expanded ensemble to sample in the same way without increase the number of coupled and uncoupled states. Wang et al. [24] included 6 coupled and 3 uncoupled replicas; factors of $f_0 = 6$ and $f_{K-1} = 3$ should achieve the same effect with state 0 being fully coupled and $K-1$ being uncoupled. We added factors $f_1 = f_{K-1} = 3$ to the simulation for 1-methylpyrrole. We intended to add these values for all ligands, but a setup script omitted them from the input files. This does not negatively affect the validity of samples in any state; it simply changes how many samples are gathered. We discuss a possible effect of this change in Sec. 4.1. We were unable to collect samples in state 0, as mentioned in the previous paragraph. Switching states more frequently can increase the rate of transition into and out of binding locations. The movement of the ligand into a binding location or back out of that location into the solvent are termed transition events, or just transitions, since the ligand transitions from bound phase to solvated phase behavior. During simulation, the ligand explores its environment with kinetics that depend on the current state. In the states closest to being fully coupled, the ligand can become kinetically trapped in the bound phase. We introduce the series of alchemical states to escape these traps by allowing the ligand to decouple from its environment. Moving to the decoupled state can take several steps, since state changes are controlled by the MG algorithm. The weights can in principle be chosen to give an equal probability of switching to any state from the fully coupled state. Choosing these weights is not a straightforward process, however, and is not something we were able to explore in detail. We noticed that many of our simulations have spans of several nanoseconds in which the states closest to fully decoupled are not visited. These nearly always correlate with the ligand visiting a binding location, particularly the experimental binding location. We hypothesize that an optimal set of weights exist that would shorten these spans and increase the number of transitions observed in a simulation.

The time needed for simulations to reach convergence can in principle be decreased by optimizing simulation parameters. Initial simulations of 1-methylpyrrole had 14 states, but poor kinetics were observed as discussed in the previous paragraph. We tested an increase in the number of states before trying to adjust the weights. We ran simulations with 18 and 24 states, where the additional states were added between existing states. Interaction parameters for the 14 and 24 states are given in Tables A.1 and A.2 respectively. These new states were placed where uncertainty in the free energy was large, and from which the state rarely switched to a more uncoupled state. These states are close to $\lambda_v = 0.25$, which Naden et al. [37] have reported as a region of increased uncertainty for soft core interaction. Both 18 and 24 states captured more transitions than simulations of 14 states. We were unable to conclude that 24 states captured more transitions than 18 states.

We tried to further increase transition by modifying the weights W_m to increase sampling in these the additional states. We did not notice an increase in the rate of transition for factors of $f = 5$ and $f = 10$ added to states 7 through 21. We removed these increased factors and used 24 states for the final simulations. Simulations with more states appear to increase the number of transitions, but they also gather physical samples at a slower rate, due to sharing more time in states other than the fully coupled state.

Results and Discussion

We present our results for the study of ligand binding to T4-lysozyme by four model ligands. Section 4.1 addresses three types of simulation to compare how efficient is each type at finding binding locations. Performance seemed to decrease in the move to expanded ensemble explicit solvent, and we determine whether the change in solvent model or in simulation method made the most impact. We then present across the remaining sections the results of the explicit solvent expanded ensemble study, and compare these results to the implicit solvent HREMD data. Binding location positions, occupancies, and free energies are the focus of these results.

We calculate the implicit solvent HREMD data presented here using simulation trajectories provided by Wang et al. [24]. They created these trajectories in the same way as the simulations used for their publication.

4.1 Sampling Performance in Binding Ensemble Simulations

Convergence time is increased due to explicit water, and not by the change to expanded ensemble. We quantify the effect of switching from implicit to explicit solvent representation on the rate of transitions from bound to solvent phases using an HREMD simulation in explicit solvent for 1-methylpyrrole. This simulation uses

Rates of Transition for Different Simulation Methods

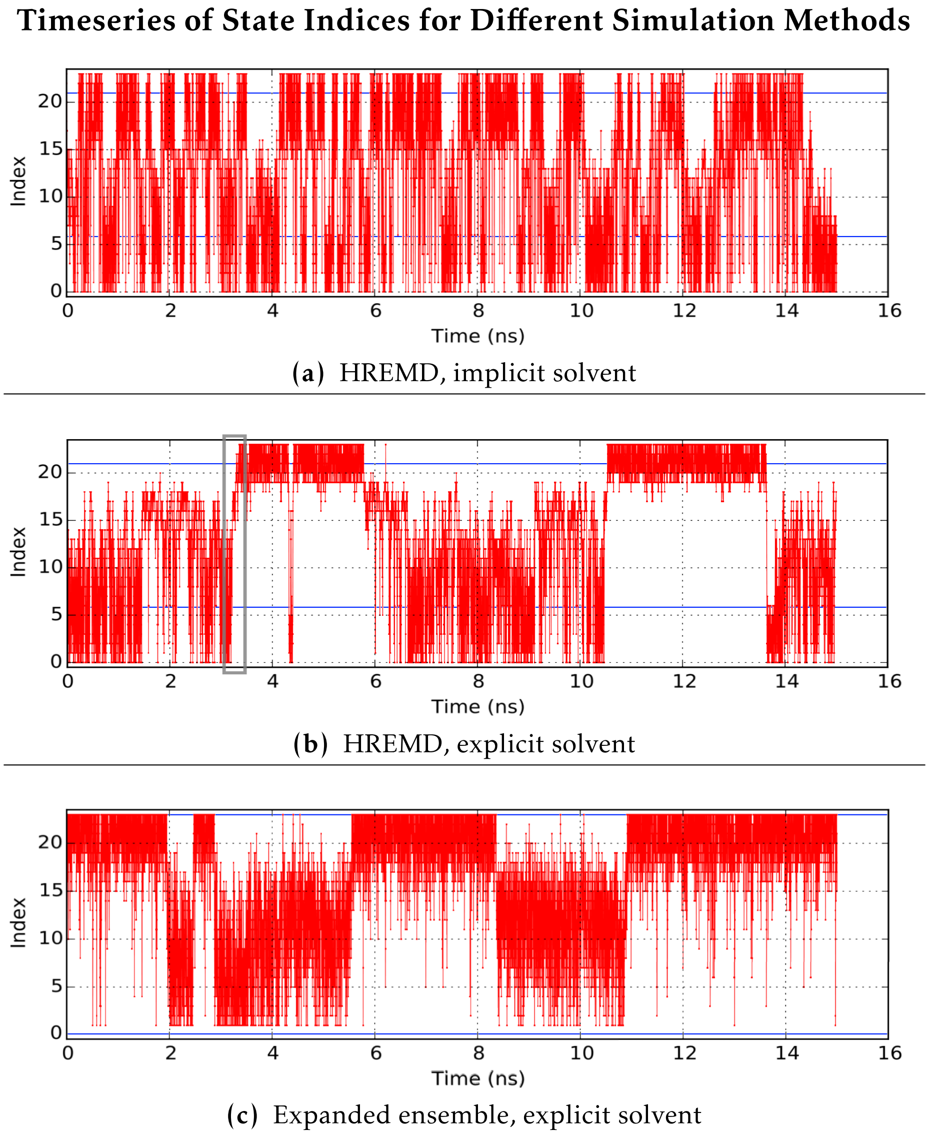
Method	HREMD	HREMD	Expanded ensemble
	implicit	explicit	explicit
benzene	10.91 ns^{-1}	-	3.73 ns^{-1}
1-methylpyrrole	10.11 ns^{-1}	0.20 ns^{-1}	0.85 ns^{-1}
para-xylene	2.71 ns^{-1}	-	1.56 ns^{-1}
phenol	4.00 ns^{-1}	-	1.72 ns^{-1}

Table 4.1: Explicit solvent representation has a greater impact on the rate of sampling than switching from HREMD to expanded ensemble. These values are calculated as an average over 360 total ns of simulation for HREMD, and 800 ns for expanded ensemble. A transition is counted whenever the simulation or replica moves to one of the fully uncoupled states after sampling in one of the fully coupled states. Values are listed as number of transitions per nanosecond, ns^{-1}. Explicit solvent HREMD was not conducted for benzene, *para*-xylene, or phenol. Higher values indicate a greater number of opportunities for the ligand to explore alternative binding locations within one nanosecond of simulation.

the same parameters as reported in the study by Wang et al. [24], but with a periodic box of solvent instead of harmonic restraints to keep the ligand near the protein. It is worth noting that Wang et al. [24] also made Monte Carlo moves to displace the ligand, which is not performed here, as it is not as useful. Proposed Monte Carlo moves would often overlap the ligand and solvent, and would almost always be rejected unless in the fully uncoupled state. Twenty-four replicas simulate 17 unique states, with 6 fully coupled and 3 fully uncoupled replicas, like the HREMD simulations of Wang et al. [24]. We measure transitions as the time it takes for a single replica, after visiting one of the fully coupled states, to sample in one of the fully uncoupled states. A total of 360 nanoseconds of simulation are collected in each of the implicit and explicit solvent HREMD simulations. This is compared to expanded ensemble explicit solvent simulations at 800 total nanoseconds, where we treat each simulation like a replica in measuring the number of transitions. As shown in Table 4.1, explicit solvent causes an order of magnitude decrease in the number of transitions per nanosecond. Expanded ensemble itself improves significantly over HREMD, sampling on average four times as many transitions. Park [23] previously offered arguments for why expanded ensemble should have higher acceptance ratios than HREMD for state switches performed between adjacent states. Our observation is consistent with their statement for "all-to-all" sampling under the metropolized-Gibbs algorithm (see Sec. 3.1). Higher acceptance ratios imply better mixing in state space, which leads to better mixing in configurational space.

We correlate the number of transitions with the ability of the simulation to find binding locations. Each time the ligand unbinds from the protein, the ligand has a chance to bind again but at a different location. We have plotted the series of state indexes versus time to compare to visual observation of the ligand trajectories for expanded ensemble explicit solvent simulation. We see a correlation between the state indexes plotted in Fig. 4.1 and the ligand leaving the protein to sample solvent. Regions where the state index remain low, where the interactions are more coupled than uncoupled, typically indicate that the ligand is sampling a binding location. We would like to see a trace closer to that of implicit solvent HREMD, where the series of state index values seem to indicate rapid binding and unbinding of the ligand. Each binding event, whether rapid or not, is still a useful sample of where the ligand spends its time.

We can theoretically predict how long a simulation will need to run to observe the same number of binding events as the implicit solvent HREMD studies. HREMD implicit solvent captures more than 10 transistions per nanosecond for two ligands, 1-methylpyrrole and benzene. Expanded ensemble explicit solvent sees less than 1 ns^{-1} for 1-methylpyrrole. This indicates that expanded ensemble simulations for 1-methylpyrrole should run for a total of ten times as many nanoseconds as the HREMD simulations to observe the same amount of transitions. The HREMD



(a) HREMD, implicit solvent

(b) HREMD, explicit solvent

(c) Expanded ensemble, explicit solvent

Figure 4.1: Explicit solvent simulations are slow to sample transitions. These simulations spend long amounts of time sampling in a narrow range of states. Plots (4.1a), (4.1b), and (4.1c) show the state index as a function of simulation time in picoseconds. Transitions can be observed visually as the span of time between the state index moving to or below the lower blue line and when it moves to or above the upper line, where the lines are drawn to show the difference between fully coupled states, intermediate states, and uncoupled states at the top. One example of a transition is the time leading up to 3.6 nanoseconds in plot (4.1b), which has been highlighted by a gray box.

implicit solvent study simulated a total of 360 ns, meaning that 3.6 μs are needed to achieve the same level of sampling different binding locations. This is just an approximation, but this quantity can be important in the design of future simulations to gauge how long a simulation should be.

Increasing the weights in the fully coupled and uncoupled states may lead to slower ligand dynamics. A sample size of three out of four ligands is small, and further study would be needed to provide convincing evidence. The trend for benzene, *para*-xylene, and phenol in Table 4.1 is that explicit solvent may be only a third or half the speed of implicit solvent, compared to a factor of 10 for 1-methylpyrrole. The only procedural differences between the three ligands and 1-methylpyrrole is that the weights for 1-methylpyrrole had been estimated from from preliminary simulations five times instead of three and that a factor of 3 was added to promote sampling in the coupled and uncoupled states. We don't attribute the slower dynamics to how many times the weights were estimated; each evaluation should give a more accurate estimate of free energies to use as weights. These improved weights lead to ligand dynamics on shorter time scales in preliminary simulations for 1-methylpyrrole. In one preliminary case, the ligand sampled the experimental site for longer than 15 ns of simulation; the adjusted weights increased the number of transitions. We instead believe that the factor of 3 caused the lower number of transitions for 1-methylpyrrole. We are unsure as to why. As a guess, the increased factor discouraged sampling in higher states in preference for the energetically favorable coupled state, and the free energy difference from coupled to uncoupled is too high to switch directly between the two. Explicit solvent HREMD for the three ligands may be able to provide more data, but we have not performed these simulations.

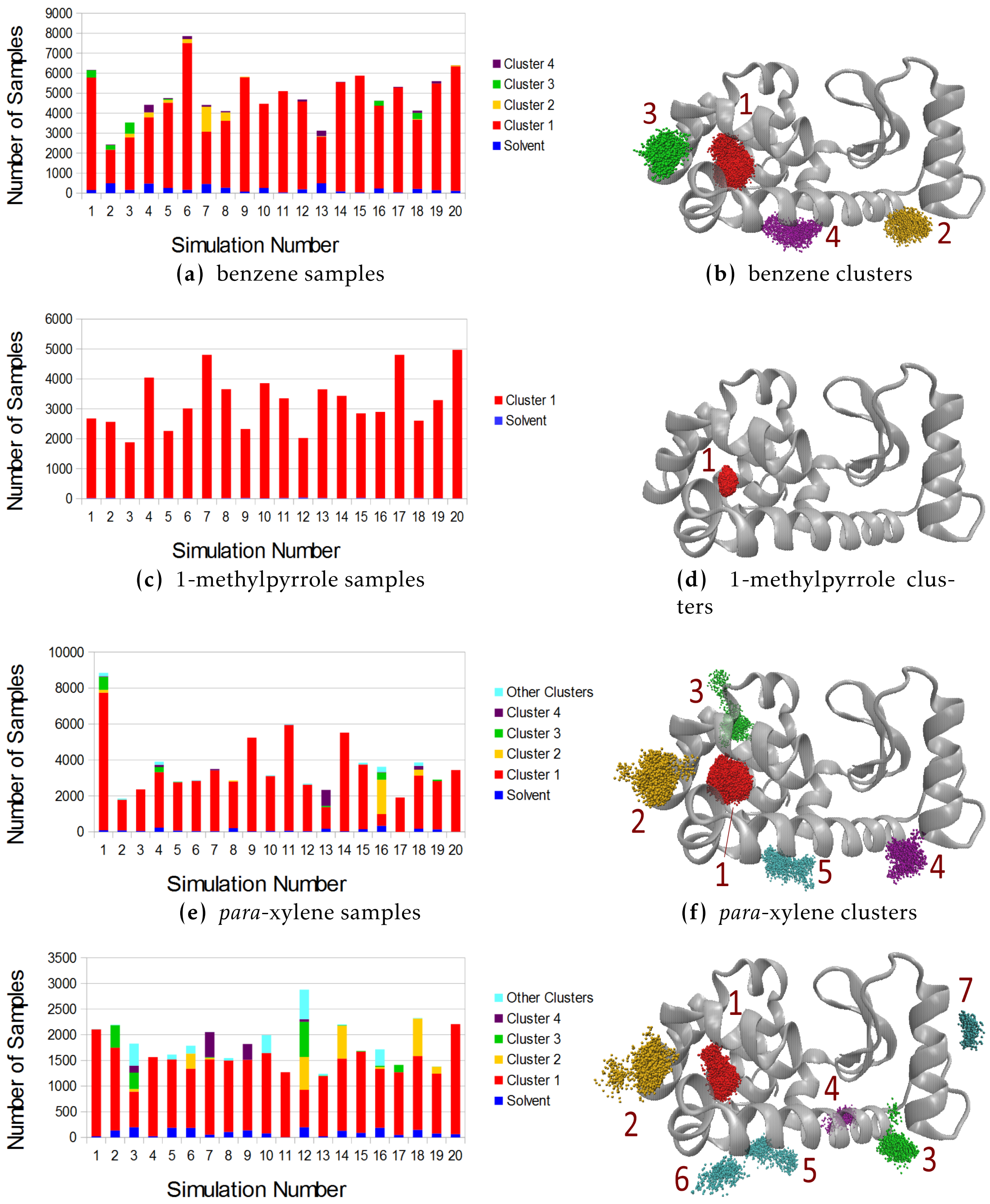
4.2 Binding Locations are Identified Using 20 Trials

We perform twenty independent simulations of T4 lysozyme L99A with each ligand. The clusters are volumes of high density of samples from alchemical state 1 which is 30% less charged than a full interacting ligand molecule. Figure 4.2 shows the number of samples for each simulation in each cluster. Samples that are outside of the protein radius but within the cutoff radius are treated as solvent samples. The protein radius is measured as the distance to the protein atom farthest from the protein center, and the cutoff radius is 5 Å past the protein radius. The density of solvent samples is used in estimating the free energy of the clusters by occupancy.

Figure 4.2 presents information that we use to judge the convergence of the binding study. Most binding locations are sampled independently by separate simulations. Sampling is still not ideal, as we would like to see each simulation visit each location for a proportionate amount of samples. However, sampling is very encouraging as each of the top 4 clusters for every ligand are visited by at least 4 simulations, with the top location visited in every simulation. The stacked bar charts show how many samples of each simulation were assigned to the clusters that are identified for each ligand. Ideal sampling would have bars in each column of approximately equal height for each cluster. Clusters are numbered by the most populous to the least. Only the four most populous clusters are shown, with any other clusters grouped together. Benzene has 4 distinct clusters, 1-methylpyrrole has 1, *para*-xylene has 5, and phenol has 7. The single cluster for 1-methylpyrrole is located in the experimental binding location. The top clusters for the other three ligands are also located in the experimental binding location.

Individual binding location free energies share better agreement with empirical occupancies for non-polar ligands. Table 4.2 show the observed values for occupancy

Number of Samples Contributed by Each Simulation to Clustering



(a) benzene samples

(b) benzene clusters

(c) 1-methylpyrrole samples

(d) 1-methylpyrrole clusters

(e) *para*-xylene samples

(f) *para*-xylene clusters

(g) phenol samples

(h) phenol clusters

Figure 4.2: We have strong confidence that binding locations which are visited in several simulations are real. Most of the identified clusters are sampled in at least 4 simulations. Details of this plot given near the beginning of Sec. 4.2

Occupancies and Free Energies by Binding Location

Location	Occupancy, O_i	Occupancy, O'_i	Free Energy, ΔG_i	Free Energy, $\Delta G'_i$
		(from ΔG_i)		(from O_i)
	%	%	kcal/mol	kcal/mol
benzene				
1	93.8 [90.5, 96.5]	91.2 [76.0, 97.1]	-3.43 ± 0.34	-3.97 [-4.20, -3.78]
2	2.7 [0.7, 5.8]	5.6 [1.7, 16.3]	-1.77 ± 0.18	-1.86 [-2.35, -1.07]
3	1.9 [0.5, 3.4]	1.5 [0.4, 5.2]	-1.00 ± 0.24	-1.64 [-2.07, -0.83]
4	1.5 [0.7, 2.6]	1.6 [0.5, 4.5]	-1.01 ± 0.15	-1.52 [-1.90, -1.03]
1-methylpyrrole				
1	100. [100., 100.]	100. [100., 100.]	-3.84 ± 0.05	- 5.40 [-5.54, -5.25]
para-xylene				
1	90.7 [84.3, 95.7]	86.0 [71.9, 92.7]	-4.03 ± 0.15	-4.29 [-4.58, -4.06]
2	3.4 [0.2, 9.4]	9.4 [3.4, 22.7]	-2.71 ± 0.29	-2.34 [-3.02, -0.29]
3	2.2 [0.4, 4.7]	1.5 [0.7, 3.1]	-1.61 ± 0.19	-2.08 [-2.61, -0.96]
4	1.9 [0.3, 5.0]	1.8 [0.7, 3.9]	-1.71 ± 0.22	-2.00 [-2.60, -0.85]
5	1.7 [0.7, 2.8]	1.3 [0.7, 2.6]	-1.55 ± 0.17	-1.91 [-2.34, -1.35]
phenol				
1	79.1 [71.8, 86.9]	7.0 [3.8, 10.7]	-1.29 ± 0.12	-3.73 [-3.94, -3.58]
2	7.4 [2.0, 13.3]	17.1 [7.4, 30.9]	-1.82 ± 0.23	-2.32 [-2.76, -1.49]
3	4.8 [0.7, 9.5]	31.2 [15.5, 48.4]	-2.18 ± 0.21	-2.06 [-2.54, -0.87]
4	2.8 [0.3, 5.8]	3.5 [0.9, 10.8]	-0.88 ± 0.37	-1.75 [-2.24, -0.40]
5	2.6 [1.0, 4.5]	28.2 [13.5, 44.7]	-2.12 ± 0.21	-1.70 [-2.07, -1.12]
6	1.7 [0.2, 3.6]	5.8 [2.3, 12.3]	-1.18 ± 0.25	-1.44 [-1.93, -0.08]
7	1.6 [0.0, 4.4]	7.1 [1.2, 31.8]	-1.30 ± 0.54	-1.41 [-2.04, 2.33]

Table 4.2: Free energy estimates agree with the number of samples in each binding location for some ligands. The binding experimental location has the highest occupancy of all identified locations, listed as location 1 for each ligand. The occupancies are calculated using the ratio of samples to the total number of clustered samples. These occupancies are used to calculate an estimate of the free energy difference $\Delta G'_i$ using Eqn. 2.15. The 95% confidence intervals for both O_i and $\Delta G'_i$ are calculated by the bootstrap method [38]. The free energy difference and uncertainty is estimated using MBAR. These free energies are used to estimate the occupancies O'_i using Eqn. 2.12. The 95% confidence intervals for O'_i are calculated assuming that MBAR estimates are normally distributed, which is deemed reasonable given the results of the study by Paliwal and Shirts [39]. The free energies $\Delta G'_i$ are within two uncertainties of ΔG_i for benzene and *para*-xylene, but larger differences are observed for 1-methylpyrrole and phenol locations. The 100% occupancy for 1-methylpyrrole is due to only one location being identified. There are other samples around the protein showing that the ligand did explore other areas, but the local density of these samples are too low to constitute a binding location.

of and the estimated binding free energies for binding locations identified by clustering. The free energies are used to estimate occupancy values, and the occupancies are compared to the number of samples in the surrounding solvent to estimate free energies. Benzene and *para*-xylene have binding free energies that are within the confidence intervals of the free energies estimated from occupancy values, showing relative agreement between free energy and occupancy. The values for 1-methylpyrrole and phenol, however, do not agree for several binding locations. The free energies estimated by occupancy are too negative for the most populated. This implies that the solvent region was not sampled often enough; a larger number of solvent samples leads to more positive free energies estimated from occupancy values. Longer simulations may lead to gathering more samples in solvent so that the ratio between the number of bound and solvent samples would agree with the binding free energy estimate. The number of samples in the experimental binding location is the highest for each ligand compared to other locations. The binding free energies do not agree with this observation for phenol, with locations 3 and 5 having the lowest free energy.

Occupancies and free energies likely do not agree because clusters were determined using undercharged samples. We were unable to collect fully coupled samples for any of the ligands. We instead used the samples in state 1 as an approximation of the binding locations. The free energies to state 1 are listed in Table 4.3. These free energies produce estimates of the occupancies that are closer to the observed occupancies for the 1-methylpyrrole and phenol. These molecules are both polar, and we expected the largest amount of error for these two, particularly for phenol. The free energy difference between states 0 and 1 for phenol is considerably larger than the other molecules, and seems to be highly dependent on the local environment. For example, location 1 has a difference in free energies of

Occupancies Based on Free Energy to Clustering State

Location	Occupancy, O_i	Occupancy, O_i'	Free Energy, ΔG_i
		(from ΔG_i to state 1)	(to state 1)
	%	%	kcal/mol
benzene			
1	93.8 [90.3, 96.5]	97.8 [93.1, 99.3]	-3.05 ± 0.34
2	2.7 [0.7, 5.9]	1.1 [0.3, 3.8]	-0.40 ± 0.18
3	1.9 [0.6, 3.7]	0.6 [0.1, 2.1]	0.02 ± 0.24
4	1.5 [0.7, 2.6]	0.5 [0.1, 1.5]	0.13 ± 0.14
1-methylpyrrole			
1	100. [100., 100.]	100. [100., 100.]	-3.29 ± 0.05
para-xylene			
1	90.7 [84.0, 95.8]	95.9 [91.1, 97.9]	-3.74 ± 0.15
2	3.4 [0.2, 9.1]	2.3 [0.8, 6.4]	-1.52 ± 0.29
3	2.2 [0.4, 4.7]	1.0 [0.4, 2.1]	-0.99 ± 0.19
4	1.9 [0.3, 4.5]	0.6 [0.2, 1.3]	-0.69 ± 0.21
5	1.7 [0.7, 2.9]	0.3 [0.1, 0.5]	-0.25 ± 0.16
phenol			
1	79.1 [72.1, 86.6]	78.4 [66.8, 85.2]	-0.01 ± 0.12
2	7.4 [2.1, 13.4]	6.0 [2.7, 12.1]	1.52 ± 0.22
3	4.8 [1.1, 9.6]	8.3 [4.1, 15.3]	1.33 ± 0.19
4	2.8 [0.2, 6.3]	0.9 [0.3, 2.9]	2.66 ± 0.35
5	2.6 [1.1, 4.5]	3.3 [1.7, 6.0]	1.88 ± 0.17
6	1.7 [0.2, 3.7]	2.0 [0.9, 4.3]	2.17 ± 0.23
7	1.6 [0.0, 4.4]	1.1 [0.2, 5.9]	2.51 ± 0.51

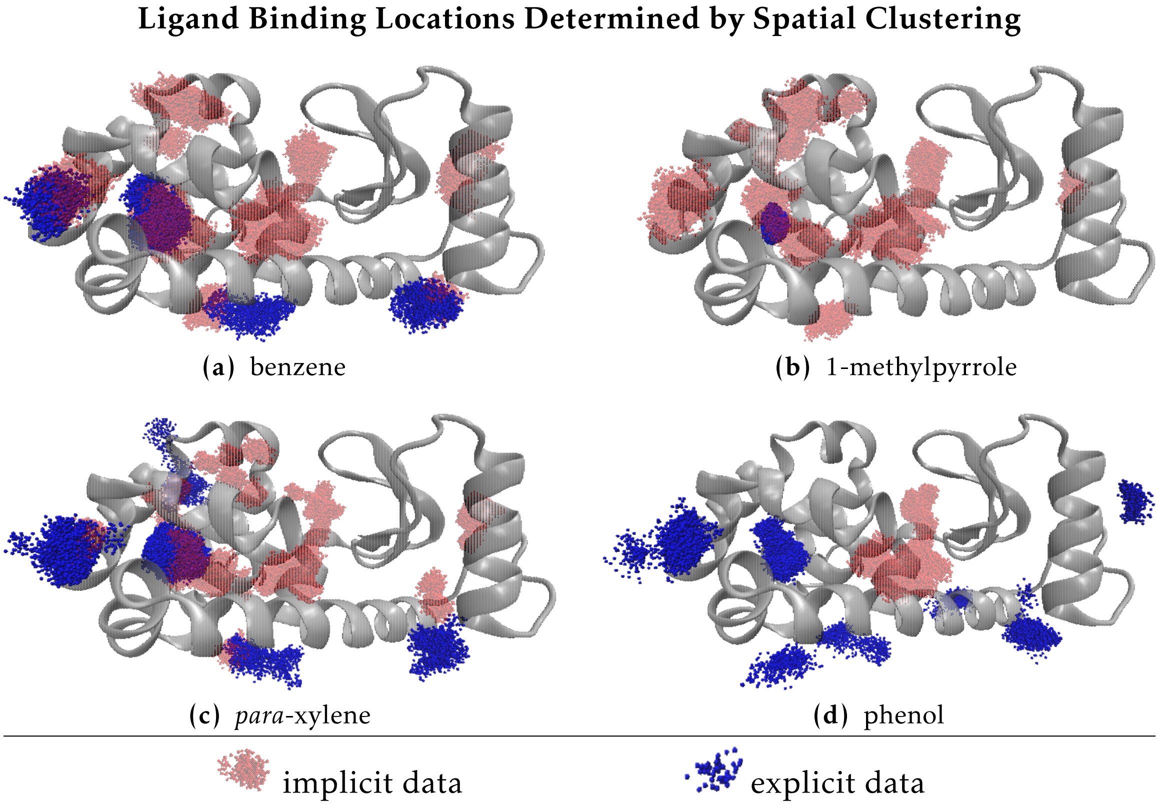
Table 4.3: Undercharged molecules lead to disagreement between free energy and occupancy estimates. The free energies listed above connect the first alchemical state, state 1, to the ligands fully interacting with only solvent. The occupancies O_i' estimated from these free energies share better agreement with the observed occupancies O_i. We interpret the disagreement between the full binding free energies and the observed occupancies to be the result of our approximation that samples in state 1 can be used to determine binding locations.

$(-1.29)-(-0.01) = -1.28$ to state 0 from state 1. Location 3 however has a difference of -3.51. This would indicate that the 30% charge difference between states 0 and 1 has the most impact in location 3, and leads to a more favorable free energy when the higher charge is present. Our approximation had a significant impact on the results, and we expect better agreement between occupancies and free energies for simulations that sample the fully coupled state.

4.3 Explicit Solvent Identifies Locations Different from Implicit Solvent Results

Simulation correctly identifies the experimental binding location as the principal binding location for three ligands. The position of samples clustered into identified locations are plotted for both implicit solvent HREMD and explicit solvent expanded ensemble results in Fig. 4.3. Most locations are not shared between the two solvation models, with phenol sharing no locations, though the experimental binding location is identified by both models for benzene, 1-methylpyrrole and *para*-xylene. Implicit solvent seems to indicate that the space between the "left" and "right" portions of the protein (as drawn) acts as a binding location for all of the ligands. Explicit shows no regions of increased sample density in that space. Explicit solvent has all four ligands visit the binding location with the highest number of samples collected there. This may be a consequence of undercharged ligand molecules, as both the experimental and computational free energies indicate that phenol does not bind with considerable affinity for the location, as determined by free energies greater than -2.47 kcal/mol [11].

We sort explicit solvent clusters into combined locations around the protein. Occupancies for each location are also sorted and shown graphically in Fig. 4.4. These combined are relabeled from (A) to (H) in decreasing total occupancy. Total occupancy is the sum of all four ligand occupancies, where zero is used for ligands that don't have a cluster at that location. The experimental binding location is location (A). Each ligand has roughly the same shape, and three have the benzene aromatic ring in common. These combined locations can reveal additional information about which functional groups are important in binding to the protein. In



(a) benzene

(b) 1-methylpyrrole

(c) *para*-xylene

(d) phenol

Figure 4.3: There are many differences between clusters determined under the two solvation models. Each point represents one ligand configuration found in areas calculated as a binding location using DBSCAN clustering. The protein is rendered in grey, with the experimental binding location in the center-left portion of each frame. Results obtained by Wang et al. [24] using HREMD in implicit solvent are plotted in a transparent red. Expanded ensemble explicit solvent results are superimposed and plotted in blue. The clusters are shown from an angle that avoids most visual overlap of clusters that are not in the same position. Locations where blue samples are overlapping red samples are locations that the implicit and explicit models share. It is difficult to know whether better agreement would be observed if fully coupled samples were collected for explicit solvent.

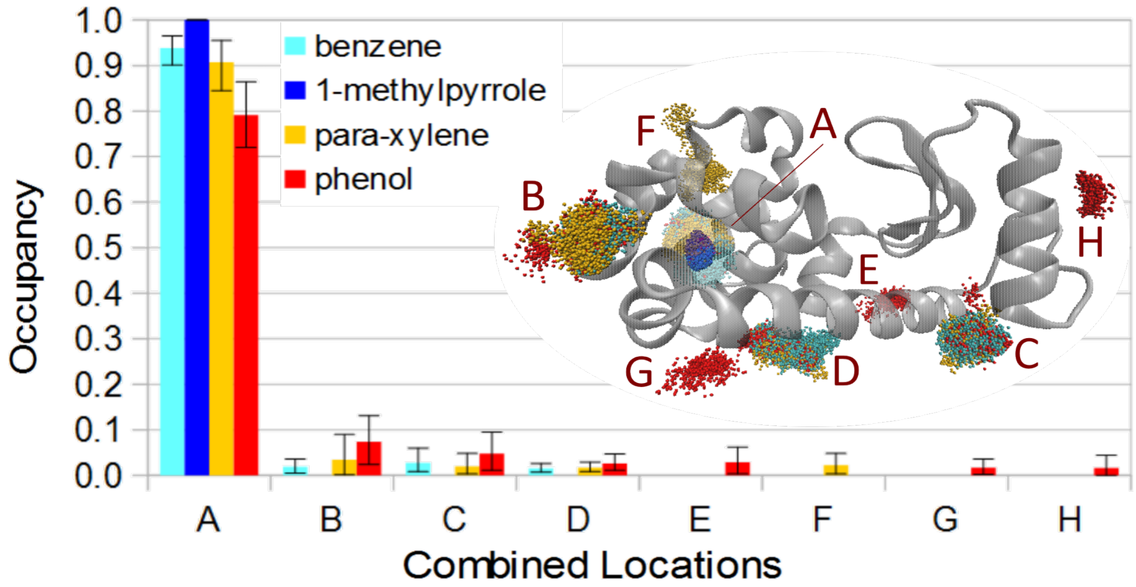
the case of lysozyme, all three benzene-based compounds bind to the same shared locations, (A)-(D). We do not make an attempt here to compare which interactions with protein residues are dominant for any of the ligands.

The specific orientations of the ligand can be plotted and potentially analyzed for important contacts with protein residues. Simulation generates orientations

Occupancies for Ligand Binding Locations in Explicit Solvent

Locations	A	B	C	D	E	F	G	H
benzene	1	3	2	4	-	-	-	-
1-methylpyrrole	1	-	-	-	-	-	-	-
para-xylene	1	2	4	5	-	3	-	-
phenol	1	2	3	5	4	-	6	7

(a) Index of clusters that belong to combined-cluster locations



(b) Occupancies for ligand locations with insert showing position of clusters

Figure 4.4: The occupancies are calculated as ratios of samples in the location relative to all the other clustered samples for all identified locations. The locations here, labeled (A) through (H), have high density of samples within the same volume relative to the protein across the four ligands. Location (A) is the experimental binding location. The table above shows which individual binding locations share the same position as other clusters for the other ligands, and gives their combined location index letter.

of the ligand within a binding location. We identify which orientations belong to each location by taking a frame from the trajectory for every point in the binding cluster. The relative position of functional groups withing the location indicates which residues are most attracted to those groups overall. Binding orientations for *para*-xylene in the experimental and one surface binding location are plotted in two figures listed in Appendix B as an example.

Free Energy Estimates for Binding to T4-lysozyme L99A

	Overall Binding Free Energy			Major Binding Location		
Ligand	ΔG_{exp}	$\Delta G_{implicit}$	$\Delta G_{explicit}$	ΔG_{Mobley}	$\Delta G_{implicit}$	$\Delta G_{explicit}$
benzene	-5.19	-5.00 ±0.01	-3.69 ±0.03	-4.56 ±0.20	-3.78 ±0.33	-3.43 ±0.34
1-meth.	-4.44	-4.96 ±0.15	-4.29 ±0.03	-4.32 ±0.08	-3.67 ±0.44	-3.84 ±0.05
p-xylene	-4.67	-6.77 ±0.02	-4.05 ±0.04	-3.54 ±0.17	-5.84 ±0.42	-4.03 ±0.15
phenol	$>$ -2.74	-5.81 ±0.03	-3.67 ±0.05	-1.26 ±0.09	-5.73 ±0.12*	-1.29 ±0.12

Table 4.4: Explicit solvent binding free energies appear to share better agreement with experiment than implicit solvent. Free energies are in kcal/mol. ΔG_{exp} is the experimentally measured value for free energy, and ΔG_{Mobley} is the estimated value for a ligand restrained to the binding location; both sets of values are reported by Mobley et al. [11]. $\Delta G_{implicit}$ are the free energies reported for each of the ligands for simulation in implicit solvent HREMD. $\Delta G_{explicit}$ are the free energies estimated here using explicit solvent expanded ensemble. Free energies to the major binding location are energies of the most populated cluster. *All major location binding free energies are for the experimental binding location, except for phenol in implicit solvent which does not have a cluster in that location.

4.4 Binding Free Energies for Both Models

Free energies agree with experiment more closely for explicit solvent than for the implicit solvent data. We estimate the binding free energy for each ligand using data from the entire simulation. The "major" binding location is the location that had the largest number of samples collected for that location. Binding free energies to the protein overall and for the major binding locations are reported in Table 4.4. The free energies to the protein overall corresponds with the experimental isothermal titration calorimetry results, as the ligands in the experiment evolve heat based on interactions to the protein as a whole. The major binding free energies correspond to the simulations that Mobley et al. [11] perform. We say that their data are for single-site simulations, as their technique samples a single localized volume within the protein. They set up a series of simulations, one in each state, but with the ligand restrained in most states so that it doesn't leave the volume of

Free Energy Estimates from Published Results

	Overall Binding Free Energy			Major Binding Location		
Ligand	ΔG_{exp}	$\Delta G_{implicit}$	$\Delta G_{explicit}$	ΔG_{Mobley}	$\Delta G_{implicit}$	$\Delta G_{explicit}$
benzene	-5.19	-6.01 ±0.81	-3.69 ±0.03	-4.56 ±0.20	-4.26 ±0.71	-3.43 ±0.34
1-meth.	-4.44	-5.05 ±0.21	-4.29 ±0.03	-4.32 ±0.08	-3.48 ±0.26	-3.84 ±0.05
p-xylene	-4.67	-5.72 ±0.95	-4.05 ±0.04	-3.54 ±0.17	-4.01 ±0.89	-4.03 ±0.15
phenol	> -2.74	-2.32 ±0.58	-3.67 ±0.05	-1.26 ±0.09	-1.03 ±0.32*	-1.29 ±0.12

Table 4.5: The published results for implicit solvent free energies share more agreement with experimental results than does explicit solvent. Free energies are in kcal/mol. Free energies listed here are in the same format as in Table 4.4, but the implicit results are from the published work of Wang et al. [24]. *All major location binding free energies are for the experimental binding location, except for phenol in implicit solvent which does not have a cluster in that location.

the experimental binding location. We use samples from our simulations that are each located within a binding location, which provides a similar dataset as what was generated for their results, but without having to know of any binding locations beforehand. All of the major binding free energies in the table are relative to the experimental binding location, except for phenol in implicit solvent.

Wang's published data [24] for implicit solvent HREMD simulation are sometimes closer to experimental and single-site binding free energies than are the explicit solvent results. These data are shown in Table 4.5. In both the published data and the trajectories analyzed here, the free energies for benzene are closer to experiment for implicit solvent. Most other explicit solvent overall and major location free energies are closer to experiment and Mobley's results respectively than implicit solvent. The accuracy of the implicit solvent estimates may be due to cancellation of error, where the assumptions involved with implicit solvent add free energy due to omitting the complex interactions with solvent, but also subtract due to ignoring the disruption of the solvent lattice that surrounds the protein. However, there is insufficient evidence here to state that explicit solvent does not also suffer from cancellation of error. Explicit solvent free energies deviate from experimental by more than 1 kcal/mol for both benzene and phenol. This may be due to the choice of force-field/assumptions in the TIP3P model for water, the lack of fully coupled samples, or assumptions in molecular dynamics itself.

Alternate binding sites contribute to the overall binding energy. We combine the free energy for all locations with the formula

$$\Delta G_{all} = -k_B T \;\ln\left(\sum_i \exp\left(\frac{-\Delta G_i}{k_B T}\right)\right), \tag{4.1}$$

so that ΔG_{all} is an estimate of binding to just the identified binding locations. The values of ΔG_{all} are -3.48 (±0.31), -3.84 (±0.05), -4.12 (±0.13), and -2.87 (±0.11) kcal/mol for benzene, 1-methylpyrrole, *para*-xylene and phenol respectively. The experimental binding location for benzene and *para*-xylene accounts for most of the free energy to the protein relative to overall binding free energy estimates for explicit solvent. There are not alternative sites for 1-methylpyrrole, but the overall free energy indicates that there is affinity to other locations around the protein, since the free energy of the major location is 0.5 kcal/mol different. Phenol has the highest contribution to free energy from alternate sites, with a difference of 1.6 kcal/mol between the all-location and major location free energies. However, the all-location energy is also 0.8 kcal/mol off of the overall free energy, indicating some favorable interaction to other areas outside the alternate binding locations.

Several issues could be addressed if the simulations are repeated. Samples should be collected in the fully coupled state, to increase confidence in the clustering results and how well the occupancies represent the free energy differences. Constant volume simulations are not a significant source of error, though the system was observed to have slightly negative pressure. The system volume can be lowered to observe if there is any effect on the predicted free energies. It is unclear if the $\ln(3)$ increase to the weights for the coupled and uncoupled states for 1-methylpyrrole was a source of error. The weights used in the Metropolized-Gibbs algorithm should be just the estimated free energies between states. The production simulations for each ligand should run without artificially increasing the weights to attempt increased sampling in the end states.

Conclusions and Future Work

We have presented a method to use computer simulation to find and evaluate binding locations of ligands to proteins. Alchemical methods allow the ligand molecule to explore more of the protein in a shorter amount of simulation time than the physical binding process would require. Additionally, alchemical states lay along a thermodynamic cycle that allows an indirect calculation of the free energy of binding to protein. Clusters of high density of samples in the fully physical state indicate where the ligand binds to the protein. We performed this study using the expanded ensemble procedure on systems of lysozyme that are explicitly solvated.

Implicit solvent studies of the same systems provided different answers to the questions of where and how much does a ligand bind. It is unclear if this difference between implicit and explicit solvation is due to the assumptions made in the implicit solvent model, or in the shortcomings of the study we perform. We demonstrated ways to evaluate how well these simulations represent the physics of the binding process, from comparing occupancies with free energies to observing the number of samples per simulation in each binding cluster. We were unable to observe free energies predicted by simulation that contain the experimental free energies within error.

It would be informative to develop ways to plot the water molecules in proximity to the protein and evaluate what interactions they have with surface binding locations. The binding location, defined as a static volume adjacent to the protein residues, should have two characteristic water densities for ligand bound and without ligand. If the density does not significantly change upon ligand binding, this suggests that the ligand is filling a hydrophobic pocket. Moderate changes in solvent density may suggest that the ligand has higher affinity for the surface residues than the water molecules, indicating that electrostatic or shape-specific interactions between ligand functional groups and the residues may be important. It would also be informative to monitor solvent molecules that appear to leave the bulk solvent, to watch them and see if they potentially have a role in mediating ligand binding to locations that otherwise would have a lower binding affinity.

This method, once validated, should be applied to other ligands and proteins. In particular, human serum albumin (HSA) plays an important role for many drugs. Establishing this method for use on HSA would provide a computational evaluation of non-specific binding, which can indicate before synthesis if a drug may have issues with pharmacokinetics. The method would also work well for study of fragment libraries, by allowing the small molecules to find as many locations as possible, potentially revealing novel target sites for specific binding. The result of the combined fragments could then be simulated as well, to confirm the predicted effect on binding.

Bibliography

(1) Mendel, C. M.; Mendel, D. B. *Biochem. J.* **1985**, *228*, 269–72.

(2) Petitpas, I.; Bhattacharya, A. a.; Twine, S.; East, M.; Curry, S. *J. Biol. Chem.* **2001**, *276*, 22804–22809.

(3) Fanali, G.; Cao, Y.; Ascenzi, P.; Trezza, V.; Rubino, T.; Parolaro, D.; Fasano, M. *IUBMB Life* **2011**, *63*, 446–451.

(4) Arepally, G. M.; Ortel, T. L. en *Annu. Rev. Med.* **2015**, *66*, 241–53.

(5) Tan, K. R.; Rudolph, U.; Lüscher, C. *Trends Neurosci.* **2011**, *34*, 188–197.

(6) Curry, S.; Brick, P.; Franks, N. P. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **1999**, *1441*, 131–140.

(7) Fasano, M.; Curry, S.; Terreno, E.; Galliano, M.; Fanali, G.; Narciso, P.; Notari, S.; Ascenzi, P. *IUBMB Life* **2005**, *57*, 787–96.

(8) Palmer, R.; Niwa, H. *Biochem. Soc. Trans.* **2003**, *31*, 973–979.

(9) Wielens, J.; Headey, S. J.; Rhodes, D. I.; Mulder, R. J.; Dolezal, O.; Deadman, J. J.; Newman, J.; Chalmers, D. K.; Parker, M. W.; Peat, T. S.; Scanlon, M. J. *J. Biomol. Screen.* **2013**, *18*, 147–159.

(10) Freire, E.; Mayorga, O. L.; Straume, M. *Anal. Chem.* **1990**, *62*, 950A–959A.

(11) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. a. *J. Mol. Biol.* **2007**, *371*, 1118–1134.

(12) Anderson, A. C. *Chem. Biol.* **2003**, *10*, 787–797.

(13) Dolezal, O.; Doughty, L.; Hattarki, M. K.; Fazio, V. J.; Caradoc-Davies, T. T.; Newman, J.; Peat, T. S. en *Aust. J. Chem.* **2013**, *66*, 1507.

(14) Blundell, T. L.; Jhoti, H.; Abell, C. *Nat. Rev. Drug Discov.* **2002**, *1*, 45–54.

(15) Bohacek, R. S.; McMartin, C. *J. Med. Chem.* **1992**, *35*, 1671–1684.

(16) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(17) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. *Curr. Comput. Aided. Drug Des.* **2011**, *7*, 146–157.

(18) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(19) Totrov, M.; Abagyan, R. *Proteins Struct. Funct. Genet.* **1997**, *29*, 215–220.

(20) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776.

(21) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

(22) Naden, L. N.; Shirts, M. R. *J. Chem. Theory Comput.* **2015**, *11*, 2536–49.

(23) Park, S. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **2008**, *77*, 016709.

(24) Wang, K.; Chodera, J. D.; Yang, Y.; Shirts, M. R. *J. Comput. Aided. Mol. Des.* **2013**, *27*, 989–1007.

(25) Okamoto, Y. *J. Mol. Graph. Model.* **2004**, *22*, 425–439.

(26) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(27) Labute, P. *J. Comput. Chem.* **2008**, *29*, 1693–8.

(28) Zhang, H.; Tan, T.; van der Spoel, D. *J. Chem. Theory Comput.* **2015**, *11*, 5103–5113.

(29) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, *12410*, 1–10.

(30) Chodera, J. D.; Shirts, M.; Beauchamp, K. A. PyMBAR., 2014, `https://github.com/choderalab/pymbar`.

(31) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047–69.

(32) Morton, A.; Baase, W. A.; Matthews, B. W. *Biochemistry* **1995**, *34*, 8564–8575.

(33) Humphrey, W; Dalke, A; Schulten, K *J. Mol. Graph.* **1996**, *14*, 33–8, 27–8.

(34) Chodera, J. D.; Shirts, M. R. *J. Chem. Phys.* **2011**, *135*, 194110.

(35) Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053.

(36) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.

(37) Naden, L. N.; Pham, T. T.; Shirts, M. R. *J. Chem. Theory Comput.* **2014**, *10*, 1128–49.

(38) Zhu, W en *Res. Q. Exerc. Sport* **1997**, *68*, 44–55.

(39) Paliwal, H.; Shirts, M. R. *J. Chem. Theory Comput.* **2011**, *7*, 4115–4134.

(40) Berendsen, H.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(41) Lindahl, E.; Hess, B.; Spoel, D. V. D. *J. Mol. Model.* **2001**, *43*, 306–317.

(42) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(43) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(44) Morton, A; Matthews, B. W. *Biochemistry* **1995**, *34*, 8576–88.

(45) Liu, L.; Baase, W. A.; Michael, M. M.; Matthews, B. W. *Biochemistry* **2009**, *48*, 8842–51.

(46) Kabsch, W. EN *Acta Crystallogr. Sect. A* **1976**, *32*, 922–923.

(47) Kabsch, W. EN *Acta Crystallogr. Sect. A* **1978**, *34*, 827–828.

(48) Bosco K Ho's Emporium - match.py - calculating the RMSD of PDB structures., `http://boscoh.com/protein/matchpy.html` (accessed 01/29/2016).

(49) Ester, M.; Kriegel, H.-p.; Sander, J.; Xu, X. In, AAAI Press: 1996, pp 226–231.

(50) Clowers, B. H. Chemometria: Density-based clustering approaches., 2008, `http://chemometria.us.edu.pl/index.php?goto=downloads` (accessed 03/20/2016).

Simulation Details

We conduct simulations using the molecular dynamics package GROMACS, version 4.6.7 [40–43]. Simulation coordinates and parameters were converted directly from those used for implicit solvent to have as direct a comparison to explicit solvent as possible. The bound configuration of 1-methylpyrrole with T4-lysozyme was converted from the files used by Mobley et al. [11] into GROMACS structure and topology files. Bound configuration or structure implies that the ligand is within the experimentally observed binding site. The parameters used for benzene, *para*-xylene, and phenol by Wang et al. [24] were converted into GROMACS topologies and use the same force-field. The structure of lysozyme/1-methylpyrrole complex was solvated, and TIP3P chosen as the water model. The bound structures of the other ligands were then created by replacing 1-methylpyrrole with the crystal-structure benzene, *para*-xylene, and phenol molecules individually, as used by Wang et al. [24]. Mobley et al. [11] confirmed the experimental binding location for 1-methylpyrrole and submitted the x-ray structure as PDB accession code 2OU0 to the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank. Morton and Matthews [44] confirmed benzene (PDB: 181L) as a binder of lysozyme, while Liu et al. [45] confirmed *para*-xylene (PDB: 3GUM).

Periodic boundary conditions applied to an explicit solvent system reduces simulation time while still providing a continuous, aqueous environment. In the periodic box, the ligand is always less than half the minimum-image distance away from the protein, in a virtual sea of water molecules. This is in contrast to implicit-solvent simulation, where a harmonic restraint was used to keep the ligand in proximity to the protein [24].

We run four different types of simulations, all with similar conditions. Solvation simulations include just the ligand in solvent; these simulations gather the samples we use to estimate solvation free energies. Randomization simulations produce unique initial configurations. Weight equilibration simulations are a series of simulations that iteratively improve the free energy estimates used as weights in the state-change algorithm. Production simulations are the sets of expanded ensemble simulations that generate orientations of the ligands which we analyze to estimate binding locations and free energies.

A.1 Solvation Simulations

Simulations of ligand in solvent do not need to use expanded ensemble since alchemical sampling is straightforward. We use 14 alchemical states for these solvation runs, Table A.1 lists the interaction parameters. The first four states discharge the ligand, while the last ten linearly decouple the Lennard-Jones term. A small periodic box of solvent surrounds the ligand in each initial configuration. These 14 configurations equilibrate at 1 atm and 300 K. Each simulation then runs for 5 nanoseconds.

A.2 Randomization Simulations

We generate structures with random ligand location to remove as much bias as possible towards the bound configuration. We start with the bound structure for

Solvation Simulations Alchemical Schedule

Index	λ_c	λ_v	Index	λ_c	λ_v	Index	λ_c	λ_v	Index	λ_c	λ_v
0	1.00	1.00	4	0.00	0.90	8	0.00	0.50	12	0.00	0.10
1	0.70	1.00	5	0.00	0.80	9	0.00	0.40	13	0.00	0.00
2	0.50	1.00	6	0.00	0.70	10	0.00	0.30			
3	0.00	1.00	7	0.00	0.60	11	0.00	0.20			

Table A.1: Since solvent is a simple environment, enhanced sampling is not necessary. Values of λ_v scale linearly from 1 to 0 with a change of 0.1 between each. State 0, with $\lambda_c = \lambda_v = 1$, is fully interacting. All charges are turned off by state 3. The Lennard-Jones term is removed entirely by state 13, representing the non-interacting state. A different set of parameters are used for randomization, weight equilibration, and production simulations, and are listed in Table A.2.

1-methylpyrrole from crystallography results, and replace this ligand with each of the other three. We use these structures as initial configurations for randomization simulations as well as for the explicit solvent HREMD study. The four simulations, one for each ligand, spend 10 nanoseconds in the non-interacting state. Each ligand then moves about the system freely, without collision with any other molecules. The protein is free to interact with the solvent and rearrange its residues according to its equilibrium ensemble. After the simulation completes, we pull frames from the trajectory at even intervals to convert into initial configuration files. The ligand may be anywhere in the simulation volume in these frames, including overlapping with solvent or protein atoms. The configurations generated for benzene are shown in Fig. A.1b.

A.3 Weight Equilibration Simulations

Weights are used in the state change algorithm to affect more even movement in state space. We use the Wang-Landau algorithm [35] as included with GROMACS in 10 nanosecond simulations to determine a first guess for the weights. We

Plots of Initial Configurations of Both Bound and Randomized Ligand Orientations

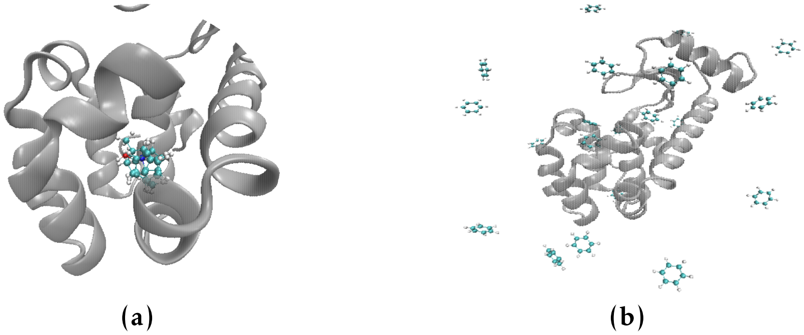


Figure A.1: Random starting locations for the ligand allows it to avoid the kinetically trapped condition of sampling the experimental binding location. This location is buried within the protein, and provides a high energetic barrier to escape. Plotted at left in (A.1a) are the ligands in the experimentally observed binding location. The protein is shown in grey, with all four ligands super-imposed. Carbon atoms are drawn in cyan, nitrogen in blue, oxygen in red, and hydrogen in white. Water molecules are not shown in these plots, but are present in the structure files. These four bound structures were used to generate randomized configurations by running decoupled simulations, with the ligand free to move throughout simulation space. Twenty initial structures are chosen from each of the four simulations to use as initial configurations in expanded ensemble production runs. The twenty initial positions for benzene are shown in (A.1b). The other ligands have similar random positions. These random initial configurations give each expanded ensemble simulation an opportunity to begin by sampling a different location than its sibling simulations. We can estimate convergence of the individual simulations by comparing the sampling in each simulation to the others. It is likely that all twenty simulations have provided sufficient sampling when each has sampled the same locations with the same frequencies.

improve upon these estimates by using MBAR on the latter 5 nanoseconds of the simulation. Another 10 nanoseconds of samples are generated without using the Wang-Landau algorithm, but instead using just the first round of MBAR free energy estimates. MBAR is applied again, this time to all 10 nanoseconds, to determine better estimates. This process could be applied iteratively until sampling is relatively even. Sampling is typically even enough when the ratio of samples in the most sampled state to the least sampled state is less than five. We only performed one more 10 nanosecond simulation, for a total of 30 nanoseconds. The production weights are determined by the estimates from MBAR for this last 10 nanoseconds.

During the Wang-Landau algorithm, a simulation starts with weights equal to zero. Every time the state might change during simulation, the Wang-Landau algorithm increases the weight of the current state by $\mathrm{d}W$. This will in turn discourage further sampling in that state so that other states will be visited. The value N_{ratio} is defined to help determine when the $\mathrm{d}W$ should change. N_{ratio} is a vector calculated as

$$N_{ratio} = \frac{N_{samples}}{\sum\limits_{l=0}^{k-1} [N_l]/k} \tag{A.1}$$

where $N_{samples}$ is the vector that counts each time a state is visited, and N_l are the values in that vector. When all values in N_{ratio} and in $1/N_{ratio}$ are greater than W_{ratio}, the number of samples are reset, and $\mathrm{d}W$ is decreased by a factor of W_{scale}. For example, assume there are three states, $k = 3$. Sampling begins in state 0, and the value $\mathrm{d}W = 1$ is added to the vector of weights.

$$\{W_0, W_1, W_2\}_t = \{1, 0, 0\}_1 \tag{A.2}$$

Suppose that simulation continues such that 20, 30, and 25 samples are collected in each state.

$$\begin{aligned} \{W_0, W_1, W_2\}_t &= \{20, 30, 25\}_{75} \\ N_{ratio} &= \{0.80, 1.25, 1.00\} \\ 1/N_{ratio} &= \{1.25, 0.80, 1.00\} \end{aligned} \tag{A.3}$$

With $W_{ratio} = 0.8$, the number of samples are reset to $\{0, 0, 0\}$ and $\mathrm{d}W = 1 * W_{scale} =$

Production Run Alchemical Schedule

Index	λ_c	λ_v	Index	λ_c	λ_v	Index	λ_c	λ_v	Index	λ_c	λ_v
0	1.00	1.00	6	0.00	0.70	12	0.00	0.43	18	0.00	0.23
1	0.70	1.00	7	0.00	0.65	13	0.00	0.40	19	0.00	0.20
2	0.50	1.00	8	0.00	0.60	14	0.00	0.37	20	0.00	0.17
3	0.00	1.00	9	0.00	0.55	15	0.00	0.33	21	0.00	0.13
4	0.00	0.90	10	0.00	0.50	16	0.00	0.30	22	0.00	0.10
5	0.00	0.80	11	0.00	0.47	17	0.00	0.27	23	0.00	0.00

Table A.2: Extra states are added in regions of high uncertainty in the free energy estimates. Naden and Shirts [22] graphically show this region of higher uncertainty near $\lambda_v = 0.25$. We observe few occurrences of simulations moving past this region to sample states with higher or lower λ_v, causing these simulations to spend 10 nanoseconds or more in just the coupled or uncoupled states. We added ten extra states at intermediate values to increase the number of transitions from smaller to higher index states. Ten was chosen to be conservative; we found that as few as four extra states may be sufficient.

0.5. Another 30, 40, 50 samples are collected, giving the following values

$$\begin{aligned} \{W_0, W_1, W_2\}_t &= \{35, 50, 50\}_{195} \\ N_{ratio} &= \{0.75, 1.00, 1.25\} \\ 1/N_{ratio} &= \{1.33, 1.00, 0.80\}. \end{aligned} \tag{A.4}$$

Sampling would continue until the values of N_{ratio} satisfy the criteria for a reset. Ideally, this process continues iteratively until dW falls below a preset tolerance, and the difference in the weights are close to the free energy differences between each state.

A.4 Production Runs

Simulations with different starting configurations should sample the same locations with sufficient run time. Twenty independent expanded ensemble simulations were performed for each ligand. Each of these simulations would be able to visit each potential binding location in the limit of infinite sampling. By starting from different initial configurations, we have more opportunities to start sampling in different binding locations. We check for consistency by noting how many simulations sampled each location.

Alignment and clustering of samples are computed by the same methods used by Wang et al. [24]. This provides a more direct comparison of the two methods. These methods assume that the density of samples are localized for a binding location. We expect the same assumption to apply in explicit solvent. Both alignment and clustering are calculated without regard for the solvent molecules; they are removed from the set of coordinates prior to analysis.

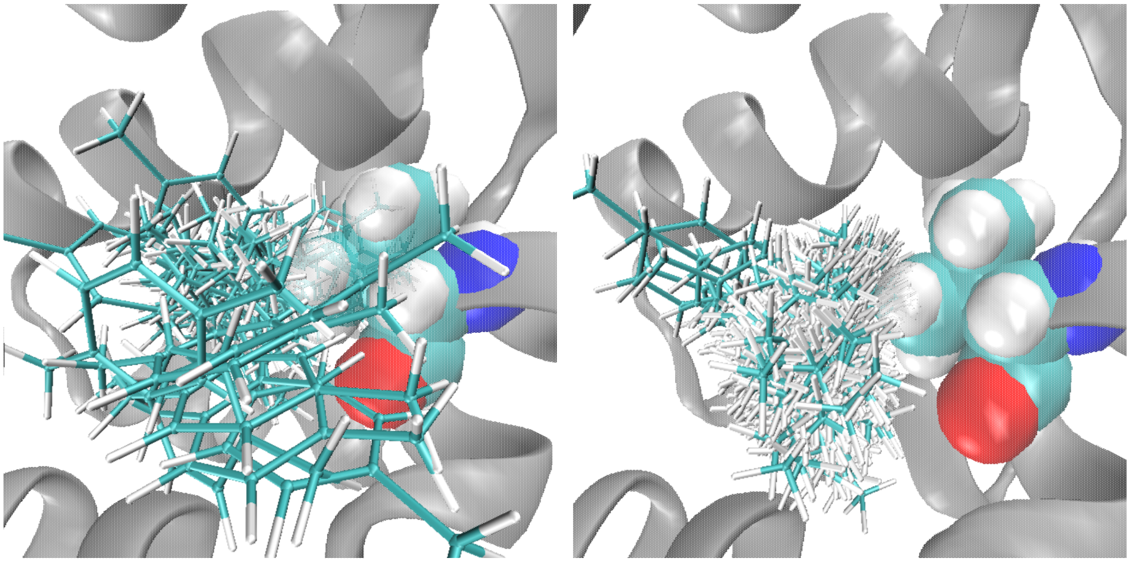
Alignment of the carbon backbone to a reference configuration is necessary to determine consistent binding locations in a flexible protein environment. During simulation, the protein is flexible, and may move, rotate and reconfigure itself according to thermal motion and interactions with itself, the ligand, and the solvent. Each frame of simulation, the protein is in a slightly different arrangement from the last frame. In order to account for this flexibility and designate a constant basis of comparison, the carbon backbone of the protein is aligned to a single configuration before clustering. A rotation-translation matrix is applied to the entire system to minimize the RMSD of the alpha-carbons between the frame and the reference configuration. This minimization is computed and applied by the Kabsch algorithm [46, 47] implemented by Ho [48]. This moves the ligand with respect to the protein, so that ligand position across multiple frames can be compared. The initial configuration of the first simulation in the set of twenty is used as the reference structure. After analysis, the clusters are aligned to the bound configuration (Fig. A.1a) in order to compare these clusters to the implicit solvent results visually.

We define a ligand binding location as a volume with increased density of samples relative to the overall sample density. Localized areas of lower free energy will be visited more often by definition of the partition function. We use the Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [49], ported into Python by Clowers [50], to automatically determine how many clusters are present, and which samples belong to those clusters. Not needing an estimate of the number of clusters reduces the information needed *a priori* about binding. Instead, the algorithm is able to isolate regions of space where the density of samples is visibly higher, using no other information than the positions of the samples themselves.

Wang et al. [24] determined that prescreening to remove low density samples improves the resolution of clusters identified by DBSCAN. We construct a grid of cubic voxels that are each 8 Å^3 and superimpose this grid over the simulation space. Voxels with a density of samples less than 2 times the average density of all voxels with at least one sample are removed. For example, there are 46,656 voxels for lysozyme which span 36 Å in every direction from the center of the protein. Assume there is at least one sample in 20,000 voxels, and there are 50,000 total samples. The "background" density would be $^5/_2$ or 2.5 samples per voxel. All samples in voxels with $2.5 * 2 = 5$ or less samples would be discarded before clustering. This helps reduce the initial level of noise and removes most samples that are visibly located within the solvent, and therefore are not useful in determining binding locations. Wang et al. [24] reported issues with rejection less than 8 times the background density, due to the close visual proximity of clusters within the experimental binding location. No issues are observed for explicit solvent using a factor of 2, as only one cluster is visually distinct in each location. In our analysis of implicit files, a factor of 8 is used for all ligands except 1-methylpyrrole, which instead used 10 to separate visibly distinct clusters.

Binding Orientations for *para*-xylene

Binding Orientations in the Experimental Binding Location

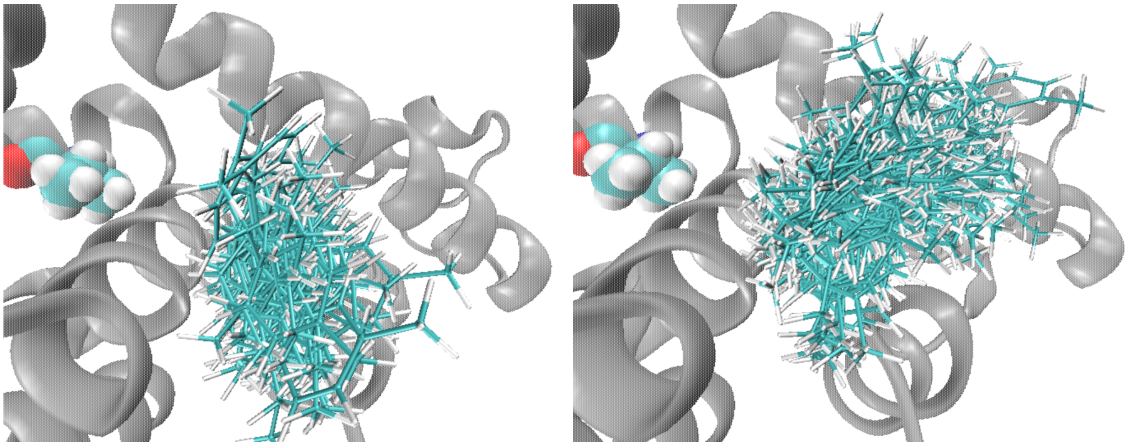


(a) *para*-xylene, implicit

(b) *para*-xylene, explicit

Figure B.1: Binding orientations of *para*-xylene are shown for both implicit and explicit solvent. The explicit solvent orientations are more compact, taking up less total volume than the collection of orientations in implicit solvent. Clusters are identified by the location of just the center-atom, highlighted in Fig. 3.1. These structures belong to the group of clustered samples, but instead of just the center atom plotted, the entire ligand is shown in stick representation.

Binding Orientations in a Surface Binding Location



(a) *para*-xylene, implicit

(b) *para*-xylene, explicit

Figure B.2: Compared with Fig. B.1, orientations in explicit solvent are spread out in a surface location as much as or more than the implicit solvent cluster. The orientation for *para*-xylene in surface location (B) are shown.

Follow-up Investigation

We performed a second set of simulations in an attempt to obtain results that should more accurately represent the binding ensemble. We addressed the issue in the code that prevented samples in the most coupled state. These repeat simulations are therefore able to sample the full range of states. We also did not increase the weights for the fully coupled and uncoupled states for 1-methylpyrrole. We were unable to redetermine the weights for each state, so that sampling is not as even across the states as it should be. These simulations also did not run as long as the original simulations, collecting roughly a third of the length for benzene, 1-methylpyrrole and *para*-xylene. A little over half the length was collected for phenol, with each of 20 simulations running for 40 ns. Overall, we expected the simulations to lack the same level of convergence as the longer, original simulations, but we do expect the identified clusters to have more physical significance as the samples are collected in the physical state.

The sections that follow present the data from the repeat simulations. These sections are analogous to the results section (Chapter 4) such that what follows could replace the original results section. However, it is an important part of this research to find that the results even without physical samples are self-consistent. We include this appendix as an addendum to the original report, so that both narratives are be recorded for use in future investigations. The data above in Chapter 4 detail the

Rates of Transition for Different Simulation Methods

Method	HREMD implicit	HREMD explicit	Expanded ensemble explicit
benzene	10.91 ns^{-1}	-	4.17 ns^{-1}
1-methylpyrrole	10.11 ns^{-1}	0.20 ns^{-1}	1.94 ns^{-1}
para-xylene	2.71 ns^{-1}	-	2.01 ns^{-1}
phenol	4.00 ns^{-1}	-	1.81 ns^{-1}

Table C.1: Explicit solvent representation has a greater impact on the rate of sampling than switching from HREMD to expanded ensemble. These values are calculated as an average over 360 total ns of simulation for HREMD, and 500 ns for expanded ensemble, 800 ns for phenol. A transition is counted whenever the simulation or replica moves to one of the fully uncoupled states after sampling in one of the fully coupled states. Values are listed as number of transitions per nanosecond, ns^{-1}. Explicit solvent HREMD was not conducted for benzene, *para*-xylene, or phenol. Higher values indicate a greater number of opportunities for the ligand to explore alternative binding locations within one nanosecond of simulation.

effects of longer, better converged simulations, while the results that follow show what is possible with physical sampling.

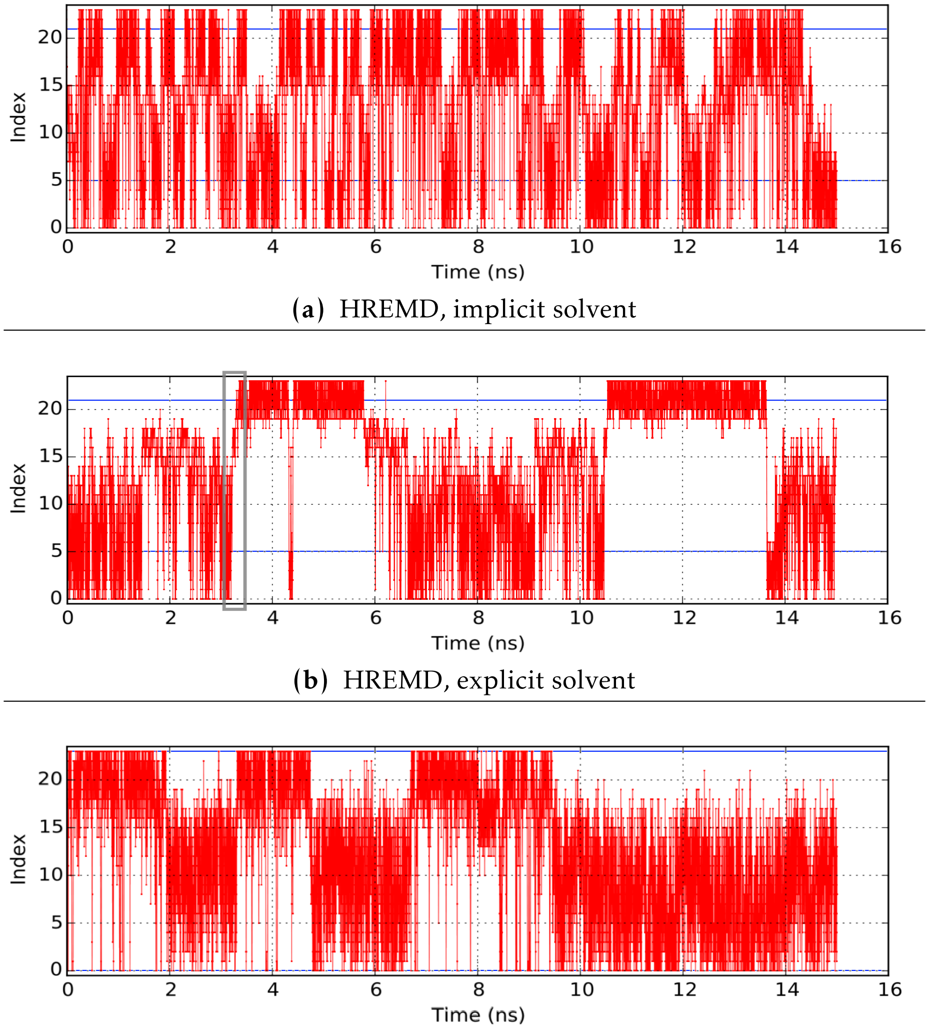
C.1 Sampling Performance in Binding Ensemble Simulations

Convergence time is increased due to explicit water, and not by the change to expanded ensemble. We quantify the effect of switching from implicit to explicit solvent representation on the rate of transitions from bound to solvent phases using a HREMD simulation in explicit solvent for 1-methylpyrrole. This simulation uses the same parameters as reported in the study by Wang et al. [24], but with a periodic box of solvent instead of harmonic restraints to keep the ligand near the protein. It is worth noting that Wang et al. [24] also made Monte Carlo moves to displace the ligand, which is not performed here, as it is not as useful. Proposed

Monte Carlo moves would often cause the ligand to overlap solvent molecules, and would almost always be rejected unless in the fully uncoupled state. Twenty-four replicas simulate 17 unique states, with 6 fully coupled and 3 fully uncoupled replicas, like the HREMD simulations of Wang et al. [24]. We measure transitions as the time it takes for a single replica, after visiting one of the fully coupled states, to sample in one of the fully uncoupled states. A total of 360 nanoseconds of simulation are collected in each of the implicit and explicit solvent HREMD simulations. This is compared to expanded ensemble explicit solvent simulations at 500 or more total nanoseconds, where we treat each simulation like a replica in measuring the number of transitions. As shown in Table C.1, explicit solvent causes two orders of magnitude decrease in the number of transitions per nanosecond. Expanded ensemble itself improves significantly over HREMD, sampling on average ten times as many transitions. Park [23] previously offered arguments for why expanded ensemble should have higher acceptance ratios than HREMD for state switches performed between adjacent states. Our observation is consistent with their statement for "all-to-all" sampling under the metropolized-Gibbs algorithm (see Sec. 3.1). Higher acceptance ratios imply better mixing in state space, which leads to better mixing in configurational space.

We correlate the number of transitions with the ability of the simulation to find binding locations. Each time the ligand unbinds from the protein, the ligand has a chance to bind again but at a different location. We have plotted the series of state indexes versus time to compare with our visual observation of the ligand trajectories for expanded ensemble explicit solvent simulation. We see a correlation between the transitions in state indexes plotted in Fig. C.1 and the ligand leaving the protein to sample solvent. Regions where the state index remain low, where the interactions are more coupled than uncoupled, typically indicate that the ligand is sampling

Timeseries of State Indices for Different Simulation Methods



(a) HREMD, implicit solvent

(b) HREMD, explicit solvent

(c) Expanded ensemble, explicit solvent

Figure C.1: Explicit solvent simulations are slow to sample transitions. These simulations spend long amounts of time sampling in a narrow range of states. Plots (C.1a), (C.1b), and (C.1c) show the state index as a function of simulation time in picoseconds. Transitions can be observed visually as the span of time between the state index moving to or below the lower blue line and when it moves to or above the upper line, where the lines are drawn to show the difference between fully coupled states, intermediate states, and uncoupled states at the top. One example of a transition is the time leading up to 3.6 nanoseconds in plot (C.1b), where the switch from coupled to uncoupled has been highlighted by a gray box.

a binding location. We would like to see a trace closer to that of implicit solvent HREMD, where the series of state index values seem to indicate rapid binding and unbinding of the ligand. Each binding event, whether rapid or not, is still a useful sample of where the ligand spends its time.

We can theoretically predict how long a simulation will need to run to observe the same number of binding events as the implicit solvent HREMD studies. HREMD implicit solvent captures more than 10 transitions per nanosecond for two ligands, 1-methylpyrrole and benzene. Expanded ensemble explicit solvent sees nearly 2 ns^{-1} for 1-methylpyrrole. This indicates that expanded ensemble simulations for 1-methylpyrrole should run for a total of five times as many nanoseconds as the HREMD simulations to observe the same amount of transitions. The HREMD implicit solvent study simulated a total of 360 ns, meaning that 1.8 μs are needed to achieve the same level of sampling different binding locations. This is just an approximation, but this quantity can be important in the design of future simulations to gauge how long a simulation should be.

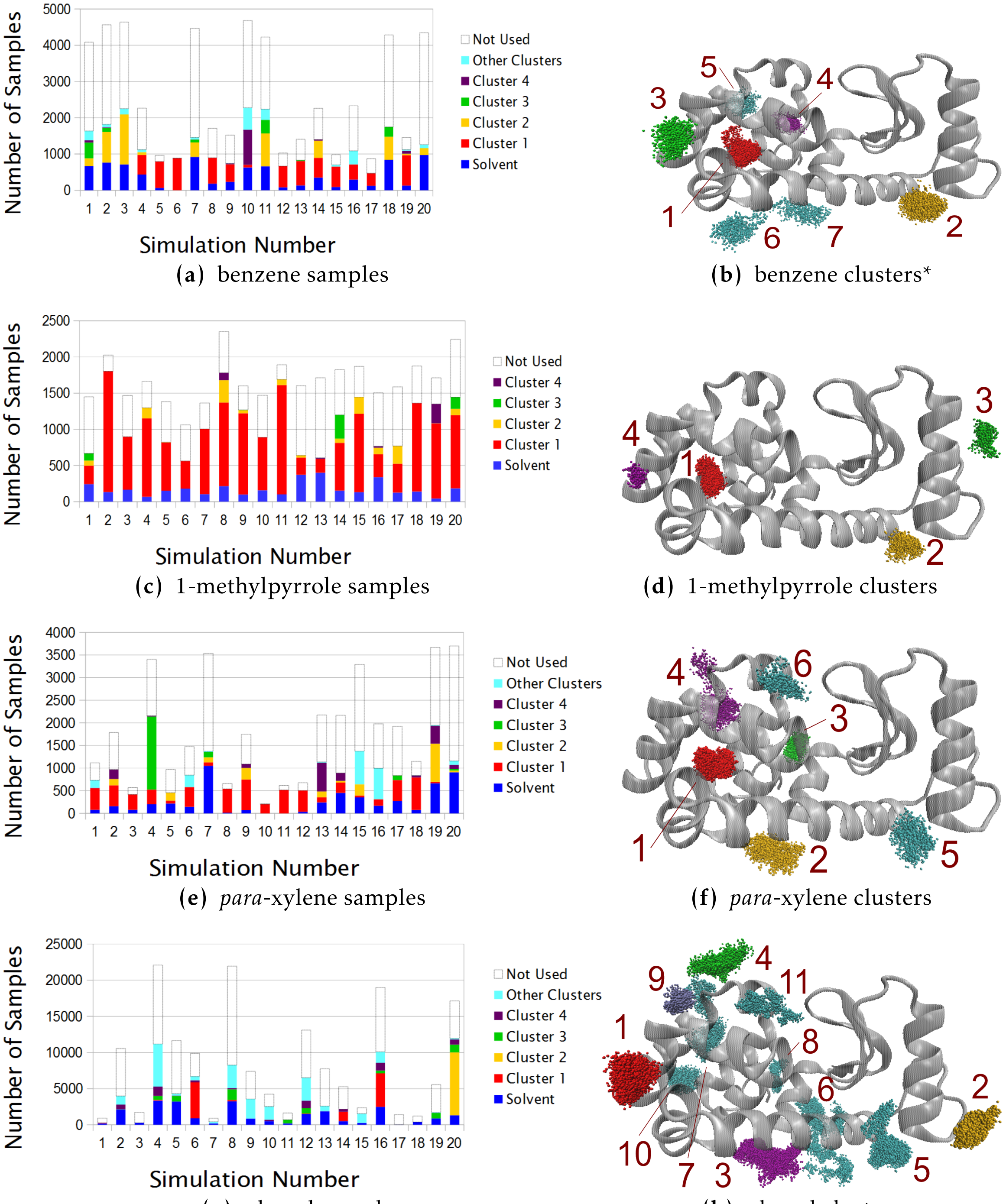
Increasing the weights in the fully coupled and uncoupled states may lead to slower ligand dynamics. Earlier simulations show that the transition rate for 1-methylpyrrole is more than twice as slow, at 0.85 ns^{-1} compared with the value observed here of 1.94 ns^{-1}. These earlier simulations added an additional factor of 3 to the weights controlling changes in state for the fully coupled and fully uncoupled states for 1-methylpyrrole. The most recent simulations reported here did not use this factor. We infer that the increased factor discouraged sampling in higher states in preference for the energetically favorable coupled state, and the free energy difference from coupled to uncoupled is too high to switch directly between the two. One out of four ligands is a small sample size, so further investigation is needed to determine how concrete is this relationship between additional factors and reduced number of transitions. Explicit solvent HREMD simulations for benzene, *para*-xylene, and phenol may also be able to provide more insight, but we have not performed these simulations.

C.2 Binding Locations are Identified Using 20 Trials

We perform twenty independent simulations of T4 lysozyme L99A with each ligand. The clusters are volumes of high density of samples from the fully interacting state 0. Figure C.2 shows the number of samples for each simulation in each cluster. Samples that are outside of the protein radius but within the cutoff radius are treated as solvent samples. The protein radius is measured as the distance to the protein atom farthest from the protein center, and the cutoff radius is 5 Å past the protein radius. The number density of solvent samples is used as a reference in estimating the free energy of the clusters by occupancy.

Figure C.2 presents information that we use to judge the convergence of the binding study. Most binding locations are sampled independently by separate simulations. Sampling is still not ideal, as we would like to see each simulation visit each location for a proportionate amount of samples. However, sampling is encouraging as each of the top 4 clusters for every ligand are visited by at least 4 simulations. For all but phenol, the most sampled location is visited by most every simulation. The stacked bar charts show how many samples of each simulation were assigned to the clusters that are identified for each ligand. Ideal sampling would have bars in each column of approximately equal height for each cluster. Clusters are numbered by the most sampled to the least. Only the four most sampled clusters are shown individually in the chart, with any other clusters grouped together. Benzene has 7 distinct clusters, 1-methylpyrrole has 4, *para*-xylene has 6, and phenol has 11.

Number of Samples Contributed by Each Simulation to Clustering



(a) benzene samples

(b) benzene clusters*

(c) 1-methylpyrrole samples

(d) 1-methylpyrrole clusters

(e) *para*-xylene samples

(f) *para*-xylene clusters

(g) phenol samples

(h) phenol clusters

Figure C.2: We have strong confidence that binding locations which are visited in several simulations are real. Most of the identified clusters are sampled in at least 4 simulations. Details of this plot given near the beginning of Sec. C.2

*Benzene's cluster 1 is located in the experimental binding location between five protein helices at the left in the images.

Individual binding location free energies do not well agree with empirical occupancies which is not unexpected for shorter simulations. Table C.2 shows the observed values for occupancy of and the estimated binding free energies for binding locations identified by clustering. The free energies are used to estimate occupancy values, and the occupancies are compared to the number of samples in the surrounding solvent to estimate free energies. All ligands have individual site binding free energies that are not within the confidence intervals of the free energies estimated from occupancy values, showing relative disagreement between free energy and occupancy. The free energies estimated by occupancy are not negative enough for nearly every location. This implies that the solvent region was sampled more often than expected given the MBAR estimates of free energy; a larger number of solvent samples leads to more positive free energies estimated from occupancy values. The occupancies do correctly predict the most sampled binding locations, as each of these has the lowest, most favorable free energy for each ligand. Longer simulations may lead to a balance between samples in solvent and samples in binding locations so that the ratio between the number of bound and solvent samples would agree with the binding free energy estimate.

Table C.2 presents several quantities to gauge the extent of sampling by the extent of agreement between free energy estimates and observed occupancies. The occupancies are used to calculate an estimate of the free energy difference $\Delta G_i'$ using Eqn. 2.15. The 95% confidence intervals for both O_i and $\Delta G_i'$ are calculated by the bootstrap method [38]. The free energy difference and uncertainty is estimated using MBAR. These free energies are used to estimate the occupancies O_i' using Eqn. 2.12. The 95% confidence intervals for O_i' are calculated assuming that MBAR

Occupancies and Free Energies by Binding Location

Location	Occupancy, O_i	Occupancy, O'_i	Free Energy, ΔG_i	Free Energy, $\Delta G'_i$
		(from ΔG_i)		(from O_i)
	%	%	kcal/mol	kcal/mol
benzene				
1	43.1 [29.4, 57.7]	83.2 [50.7, 95.5]	-3.77 ±0.44	-2.11 [-2.41, -1.80]
2	30.1 [14.6, 45.8]	8.5 [2.1, 25.5]	-2.41 ±0.17	-1.90 [-2.31, -1.32]
3	7.6 [1.8, 15.4]	1.8 [0.4, 6.7]	-1.50 ±0.25	-1.07 [-1.56, -0.06]
4	7.0 [0.5, 18.4]	1.1 [0.2, 4.2]	-1.18 ±0.30	-1.02 [-1.71, 0.56]
5	4.7 [1.0, 9.7]	1.8 [0.4, 7.2]	-1.47 ±0.30	-0.78 [-1.27, 0.22]
6	3.8 [1.3, 7.5]	2.3 [0.5, 8.2]	-1.63 ±0.25	-0.67 [-1.10, -0.00]
7	3.7 [1.3, 7.1]	1.4 [0.3, 4.7]	-1.32 ±0.23	-0.64 [-1.04, -0.05]
1-methylpyrrole				
1	87.3 [81.1, 92.7]	94.0 [83.6, 96.9]	-3.87 ±0.09	-3.11 [-3.31, -2.91]
2	7.4 [3.6, 11.7]	3.1 [1.4, 6.2]	-1.83 ±0.21	-1.64 [-1.96, -1.20]
3	3.2 [0.1, 7.5]	2.6 [0.5, 12.2]	-1.72 ±0.50	-1.14 [-1.67, 1.33]
4	2.2 [0.1, 5.7]	0.3 [0.0, 2.6]	-0.50 ±0.64	-0.91 [-1.51, 1.09]
para-xylene				
1	45.8 [32.1, 64.4]	65.2 [38.5, 83.0]	-3.32 ±0.28	-2.32 [-2.71, -1.96]
2	14.3 [4.9, 26.1]	10.5 [4.2, 20.2]	-2.23 ±0.17	-1.63 [-2.15, -0.88]
3	13.6 [0.4, 32.4]	4.8 [1.3, 14.8]	-1.77 ±0.35	-1.59 [-2.35, 0.66]
4	12.0 [3.7, 23.2]	12.8 [4.5, 28.9]	-2.35 ±0.25	-1.52 [-2.06, -0.68]
5	8.0 [0.2, 19.9]	2.6 [0.7, 8.5]	-1.40 ±0.35	-1.27 [-1.92, 1.09]
6	6.4 [0.1, 16.7]	4.0 [1.1, 12.7]	-1.66 ±0.35	-1.14 [-1.81, 1.22]
phenol				
1	20.6 [1.4, 41.4]	27.4 [2.0, 45.3]	-2.37 ±0.27	-1.74 [-2.33, -0.04]
2	15.9 [0.0, 40.7]	0.5 [0.0, 76.7]	-0.01 ±2.03	-1.59 [-2.34, inf]
3	11.3 [5.3, 22.4]	22.4 [1.6, 48.5]	-2.25 ±0.41	-1.38 [-1.77, -0.86]
4	10.8 [4.8, 20.2]	11.3 [0.9, 17.9]	-1.84 ±0.17	-1.35 [-1.73, -0.84]
5	9.7 [0.5, 21.0]	9.2 [0.7, 18.4]	-1.72 ±0.27	-1.29 [-1.84, 0.44]
6	9.5 [2.0, 24.5]	10.4 [0.8, 20.8]	-1.79 ±0.27	-1.28 [-1.86, -0.24]
7	6.9 [0.8, 16.7]	6.0 [0.5, 13.6]	-1.46 ±0.31	-1.09 [-1.62, 0.15]
8	4.1 [0.0, 13.8]	2.9 [0.1, 20.8]	-1.04 ±0.73	-0.78 [-1.53, 3.52]
9	3.9 [0.1, 9.1]	7.3 [0.1, 83.5]	-1.58 ±1.31	-0.75 [-1.27, 1.36]
10	3.9 [0.4, 8.9]	1.4 [0.1, 4.0]	-0.58 ±0.39	-0.74 [-1.24, 0.74]
11	3.4 [0.0, 11.7]	1.3 [0.1, 5.6]	-0.53 ±0.53	-0.66 [-1.40, inf]

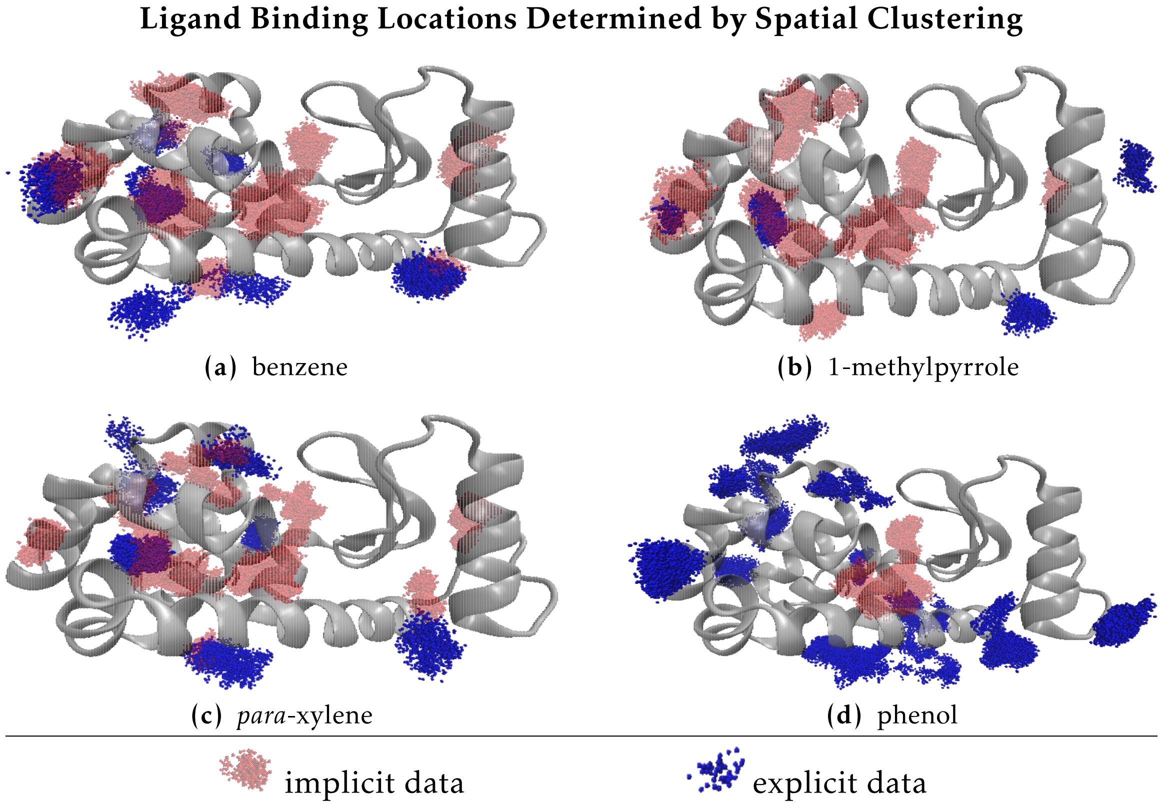
Table C.2: Free energy estimates agree with the number of samples in each binding location for some ligands. The binding experimental location has the highest occupancy of all identified locations, listed as location 1 for each ligand. The occupancies are calculated using the ratio of samples to the total number of clustered samples. Explanation of these values is given at the end of Section C.2.

estimates are normally distributed, which is deemed reasonable given the results of the study by Paliwal and Shirts [39]. We note that these short simulations were only able to produce qualitative agreement between occupancies and free energies, with the most-favorable free energies found for the most-sampled locations.

C.3 Explicit Solvent Identifies Locations Different from Implicit Solvent Results

Simulation correctly identifies the experimental binding location as the principal binding location for three ligands. The position of samples clustered into identified locations are plotted for both implicit solvent HREMD and explicit solvent expanded ensemble results in Fig. C.3. Most locations are not shared between the two solvation models, though the experimental binding location is identified by both models for benzene, 1-methylpyrrole and *para*-xylene. Implicit solvent seems to indicate that the space between the “left” and “right” portions of the protein (as drawn) acts as a binding location for all of the ligands. Explicit shows no regions of increased sample density in that space. Both implicit and explicit do not find a binding cluster in the experimental binding location for phenol, which is expected for this non-binder.

We sort explicit solvent clusters into combined locations around the protein. Occupancies for each location are also sorted and shown graphically in Fig. C.4. These combined locations are relabeled from (A) to (F) in decreasing total occupancy. Total occupancy is the sum of all four ligand occupancies, where zero is used for ligands that don’t have a cluster at that location. The experimental binding location is location (A). Each ligand has roughly the same shape, and three have the benzene aromatic ring in common. These combined locations can reveal additional information about which functional groups are important in binding to the protein.



(a) benzene

(b) 1-methylpyrrole

(c) *para*-xylene

(d) phenol

Figure C.3: There are many differences between clusters determined under the two solvation models. Each point represents one ligand configuration found in areas calculated as a binding location using DBSCAN clustering. The protein is rendered in grey, with the experimental binding location in the center-left portion of each frame. Results obtained by Wang et al. [24] using HREMD in implicit solvent are plotted in a transparent red. Expanded ensemble explicit solvent results are superimposed and plotted in blue. The clusters are shown from an angle that avoids most visual overlap of clusters that are not in the same position. Locations where blue samples are overlapping red samples are typically locations that the implicit and explicit models share.

We do not make an attempt here to compare which interactions with protein residues are dominant for any of the ligands.

The specific orientations of the ligand can be plotted and potentially analyzed for important contacts with protein residues. Simulation generates orientations of the ligand within a binding location. We identify which orientations belong to each location by taking a frame from the trajectory that corresponds to each

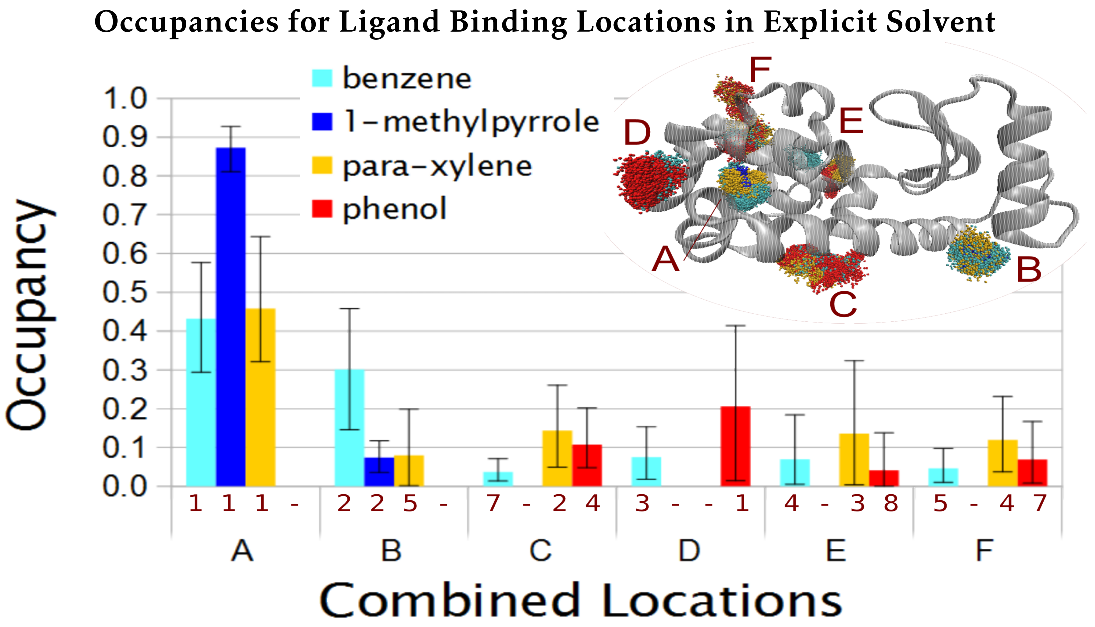


Figure C.4: The occupancies are calculated as ratios of samples in the location relative to all the other clustered samples for all identified locations. The locations here, labeled (A) through (F), have high density of samples within the same volume relative to the protein across the four ligands. Location (A) is the experimental binding location. The individual binding location numbers are given in maroon just above the combined location index letter in the chart. The combined locations are sorted by total occupancy across all four ligands. Error bars represent 95% confidence intervals in the individual binding location occupancies.

point in the binding cluster. The relative position of functional groups within the location indicates which residues are most attracted to those groups overall. Binding orientations for *para*-xylene in the experimental and one surface binding location as determined by earlier simulations that did not sample the fully coupled state are plotted in two figures listed in Appendix B as an example.

C.4 Binding Free Energies for Both Models

Free energies agree with experiment somewhat more closely for explicit solvent than for the implicit solvent data. We estimate the binding free energy for each

Free Energy Estimates for Binding to T4-lysozyme L99A

	Overall Binding Free Energy			Major Binding Location		
Ligand	ΔG_{exp}	$\Delta G_{implicit}$	$\Delta G_{explicit}$	ΔG_{Mobley}	$\Delta G_{implicit}$	$\Delta G_{explicit}$
benzene	-5.19	-5.00 ±0.01	-3.56 ±0.03	-4.56 ±0.20	-3.78 ±0.33	-3.77 ±0.44
1-meth.	-4.44	-4.96 ±0.15	-3.94 ±0.03	-4.32 ±0.08	-3.67 ±0.44	-3.87 ±0.09
p-xylene	-4.67	-6.77 ±0.02	-3.86 ±0.04	-3.54 ±0.17	-5.84 ±0.42	-3.32 ±0.28
phenol	$>$ -2.74	-5.81 ±0.03	-3.61 ±0.03	-1.26 ±0.09	-5.73 ±0.12*	-2.37 ±0.27*
	Close Binding Free Energy			All Location Free Energy		
Ligand		$\Delta G_{implicit}$	$\Delta G_{explicit}$		$\Delta G_{implicit}$	$\Delta G_{explicit}$
benzene		-4.72 ±0.01	-3.44 ±0.05		-4.58 ±0.11	-3.89 ±0.36
1-meth.		-4.96 ±0.15	-3.88 ±0.12		-4.54 ±0.14	-3.92 ±0.09
p-xylene		-6.77 ±0.03	-3.75 ±0.08		-6.19 ±0.25	-3.57 ±0.19
phenol		-5.81 ±0.03	-3.50 ±0.07		-5.73 ±0.12	-3.14 ±0.16

Table C.3: Explicit solvent binding free energies appear to share better agreement with experiment than implicit solvent, though not by much. Free energies are in kcal/mol. ΔG_{exp} is the experimentally measured value for free energy, and ΔG_{Mobley} is the estimated value for a ligand restrained to the binding location; both sets of values are reported by Mobley et al. [11]. $\Delta G_{implicit}$ are the free energies estimated here for each of the ligands for simulation in implicit solvent HREMD. $\Delta G_{explicit}$ are the free energies estimated here using explicit solvent expanded ensemble. Free energies to the major binding location are energies of the most populated cluster. *All major location binding free energies are for the experimental binding location, except for phenol in both implicit and explicit solvent which do not have a cluster in that location. "Close binding" free energies include a solvent and volume correction to reduce any over-estimation of the free energy by including favorable interactions to the "solvent volume" portion of the simulation space. "All location" free energies are explained at the end of Section C.4, and in general show how much free energy would be ignored if only binding to a single location is considered.

ligand using data from the entire simulation. The "major" binding location is the location that had the largest number of samples collected for that location. A close binding free energy adds a small correction to account for possible overestimation due to the size of the simulation box, and an all location free energy includes contributions from all identified binding locations. All four of these definitions of binding free energies are reported in Table C.3. The free energies to the protein overall correspond with the experimental isothermal titration calorimetry results,

Free Energy Estimates from Published Results

	Overall Binding Free Energy			Major Binding Location		
Ligand	ΔG_{exp}	$\Delta G_{implicit}$	$\Delta G_{explicit}$	ΔG_{Mobley}	$\Delta G_{implicit}$	$\Delta G_{explicit}$
benzene	-5.19	-6.01 ±0.81	-3.56 ±0.03	-4.56 ±0.20	-4.26 ±0.71	-3.77 ±0.44
1-meth.	-4.44	-5.05 ±0.21	-3.94 ±0.03	-4.32 ±0.08	-3.48 ±0.26	-3.87 ±0.09
p-xylene	-4.67	-5.72 ±0.95	-3.86 ±0.04	-3.54 ±0.17	-4.01 ±0.89	-3.32 ±0.28
phenol	> -2.74	-2.32 ±0.58	-3.61 ±0.03	-1.26 ±0.09	-1.03 ±0.32*	-2.37 ±0.27*

Table C.4: The published results for implicit solvent free energies share more agreement with experimental results than does explicit solvent. Free energies are in kcal/mol. Free energies listed here are in the same format as in Table C.3, but the implicit results are from the published work of Wang et al. [24]. *All major location binding free energies are for the experimental binding location, except for phenol in implicit and explicit solvent which do not have a cluster in that location.

as the ligands in the experiment evolve heat based on interactions to the protein as a whole. The major binding free energies correspond to the simulations that Mobley et al. [11] perform. We say that their data are from single-site simulations, as their technique samples a single localized volume within the protein. They set up a series of simulations, one in each state, but with the ligand restrained in most states so that it doesn't leave the volume of the experimental binding location. We use samples from our simulations that are each located within a binding location, which provides a similar dataset as what was generated for their results, but without having to know of any binding locations beforehand. All of the major binding free energies in the table are relative to the experimental binding location, except for phenol in implicit and explicit solvent.

Wang's published data [24] for implicit solvent HREMD simulation are sometimes closer to experimental and single-site binding free energies than are the explicit solvent results. These data are shown in Table C.4. In both the published data and the trajectories analyzed here, the free energies for benzene and 1-methylpyrrole are closer to experiment for implicit solvent. Some other explicit solvent overall and major location free energies are closer to experiment and Mobley's results respectively than implicit solvent. The accuracy of the implicit solvent estimates may be due to cancellation of error, where the assumptions involved with implicit solvent add free energy due to omitting the complex interactions with solvent, but also subtract due to ignoring the disruption of the solvent lattice that surrounds the protein. However, there is insufficient evidence here to state that explicit solvent does not also suffer from cancellation of error. Explicit solvent free energies deviate from experimental by more than 1 kcal/mol for both benzene and phenol. This may be due to the choice of force-field/assumptions in the TIP3P model for water, or assumptions in molecular dynamics itself.

Alternate binding sites contribute to the overall binding energy. We combine the free energy for all locations with the formula

$$\Delta G_{all} = -k_B T \ \ln\left(\sum_i \exp\left(\frac{-\Delta G_i}{k_B T}\right)\right), \tag{C.1}$$

so that ΔG_{all} is an estimate of binding to just the identified binding locations. The values of ΔG_{all} are -3.89 (±0.36), -3.92 (±0.09), -3.57 (±0.19), and -3.14 (±0.16) kcal/mol for benzene, 1-methylpyrrole, *para*-xylene and phenol respectively. These values are listed in Table C.3.